پاسخ‌دهی خودکار به پرسش‌های مربوط به محتوای تصاویر به زبان فارسی با استفاده از تکنیک‌های مبتنی بر یادگیری عمیق

امیر شکری [۱]، علیرضا غلام‌نیا [۲]

[۱] دانش‌آموخته کارشناسی‌ارشد هوش‌مصنوعی، دانشگاه سمنان – amirsh.nll@gmail.com

[۲] دانشجو کارشناسی‌ارشد هوش‌مصنوعی، دانشگاه سمنان - Gholamniareza@gmail.com

چکیده

امروزه پاسخ‌دهی خودکار به پرسش‌های مربوط به محتوای تصاویر (سیستم پرسش و پاسخ تصویری) کاربرد فراوانی دارد. در سیستم‌های پرسش و پاسخ تصویری، یک تصویر و یک سوال متنی در مورد تصویر به عنوان ورودی در نظر گرفته می‌شود و این سیستم باید پاسخ صحیح به پرسش مطرح شده را پیش‌بینی کند. هدف اصلی در این سیستم‌ها بالا بودن دقت صحت پاسخ پیش‌بینی‌شده است. برای این منظور عوامل مختلفی از جمله انتخاب شبکه‌های عصبی مناسب جهت پردازش ورودی‌ها و انتخاب مجموعه‌داده مناسب بسیار مهم است. همچنین استفاده از انواع مختلف سازوکار توجه در مدل می‌تواند باعث بهبود عملکرد کلی سیستم پرسش و پاسخ تصویری شود. تا به امروز پژوهش‌های اندکی در مورد سیستم‌های پرسش و پاسخ تصویری به زبان فارسی انجام شده است. از همین رو در این مقاله به معرفی یک سیستم پرسش و پاسخ تصویری به زبان فارسی پرداختیم. در مدل پیشنهادی، ما از شبکه عصبی کانولوشنی با معماری ResNext جهت پردازش تصویر استفاده کردیم که برای اولین بار در سیستم پرسش و پاسخ تصویری استفاده شده است. برای پردازش متن ورودی نیز از شبکه عصبی بازگشتی از نوع حافظه کوتاه مدت طولانی دوسویه استفاده کردیم. همچنین از دو نوع سازوکار توجه در مدل پیشنهادی استفاده شده است. نتیجه حاصل شده نشان می‌دهد که دقت صحت پاسخ پیش‌بینی شده در مدل پیشنهادی این مقاله، بالاترین مقدار بدست آمده نسبت به نمونه‌های موجود به زبان فارسی است. جزئیات پیاده سازی و کدهای این مقاله در لینک زیر موجود می‌باشد:

https://github.com/amirshnll/persian-visual-question-answering

واژگان کلیدی: سیستم پرسش و پاسخ تصویری، شبکه عصبی کانولوشنی، شبکه عصبی بازگشتی، سازوکار توجه

The 18th National Conference
25/02/2023-(Mazandaran)
Elmavaran Danesh Group
R.S. Institute
Article Code: NCCIT -99030

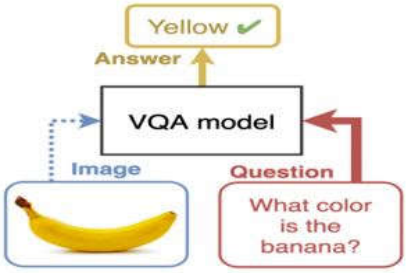
Indexing Accepted Articles in ***Civilica***

۱- مقدمه

یکی از حوزه‌های مورد علاقه پژوهشگران و محققان جهت پژوهش در زمینه بینایی کامپیوتر[1] و پردازش زبان طبیعی[2] مبحث میان‌رشته‌ای پرسش و پاسخ تصویری[3] است. در پرسش و پاسخ تصویری پیش‌بینی پاسخ صحیح به پرسش مطرح شده در مورد تصویر، هدف اصلی است. در این مسئله ماشین باید بتواند تصویر و متن را درک کرده و بین آن‌ها ارتباط برقرار کند. یکی از چالش‌های جالب در این حوزه، چگونگی ادغام و ترکیب پردازش زبان طبیعی و بینایی کامپیوتر جهت رسیدن به پاسخ مناسب است [۲].

شبکه‌ی عصبی و مجموعه‌داده، نقش بسیار مهم و کلیدی در سیستم پرسش و پاسخ تصویری دارند. در این سامانه‌ها از شبکه‌های عصبی مختلف استفاده می‌شود تا بهترین نتیجه حاصل شود. با پیشرفت‌هایی که در حوزه سخت‌افزار در کنار حوزه نرم‌افزار حاصل شده است، مجموعه‌داده‌های بزرگ و مناسبی در زمینه‌ی پرسش و پاسخ تصویری تولید شده است، که استفاده از آن‌ها باعث بهبود چشمگیری در نتیجه نهایی می‌شود. به دلیل کارایی فراوان سیستم پرسش و پاسخ تصویری در عرصه‌های مختلف نظیر پزشکی، صنعتی و ... پژوهشگران در تلاش هستند که با استفاده از شبکه‌های عصبی کاراتر و جدیدتر و سازوکارهای مختلف نظیر سازوکار توجه[4] در این سامانه‌ها باعث بهبود حداکثری در پیش‌بینی صحیح پاسخ شوند.

در سیستم پرسش و پاسخ تصویری، چالش‌ها و مشکلاتی وجود دارد که باید در نظر گرفته شوند. از جمله این موارد می‌توان به نبود مجموعه‌داده مناسب فارسی، نیاز به سخت‌افزار قوی، نحوه بدست آوردن بازنمایی مشترک به وسیله‌ی ترکیب ویژگی‌های بدست آمده از متن و تصویر و ... اشاره کرد. با مطالعات پژوهش‌های مربوط به پرسش و پاسخ تصویری، نکته‌ای که کاملا مشهود است، تاثیر چشمگیر استفاده از انواع سازوکار توجه در کارایی و دقت پیش‌بینی پاسخ در این سیستم‌ها است. در شکل ۱، شماتیک کلی یک مدل پرسش و پاسخ تصویری نشان داده شده است که یک تصویر و یک پرسش را به عنوان ورودی دریافت می‌کند و خروجی مناسب (پاسخ برای پرسش) را پیش بینی می‌کند.



شکل ۱: ساختار کلی سیستم پرسش و پاسخ تصویری

[1] Computer Vision
[2] Natural Language Processing (NLP)
[3] Visual Question Answering (VQA)
[4] Attention Mechanism

هدف این پژوهش طراحی یک سامانه پرسش و پاسخ تصویری است که دو ورودی (یک تصویر و یک پرسش مرتبط با تصویر) به زبان فارسی را دریافت کرده و قادر باشد به پرسش مطرح شده از محتوای تصویر به زبان فارسی با دقت قابل قبولی پاسخ دهد. سامانه طراحی شده در این پژوهش می‌تواند در زمینه‌های مختلف کارایی فراوانی داشته باشد و استفاده‌های فراوانی از آن شود.

۲- بیان مسئله

امروزه جهان با سرعت زیادی به سمت ماشینی شدن به پیش می‌رود. بسیاری از امور در حال حاضر با ماشین‌ها و ربات‌ها انجام می‌شود و روز به روز نقش انسان‌ها در کارها کمتر شده و جای خود را به ماشین‌های هوشمند می‌دهند. به طور کلی وجود ماشین‌ها و سامانه‌های هوشمند باعث راحت شدن زندگی برای بشر می‌شود. تعامل انسان و کامپیوتر در عصر حاضر، وارد مرحله‌ی جدیدی شده است. در بسیاری از امور مختلف در جهان امروز کامپیوترها توانسته‌اند به انسان‌ها کمک فراوانی کنند. با توجه به پیشرفت‌هایی که در هوش‌مصنوعی و یادگیری عمیق به وجود آمده و علاقه پژوهشگران برای پژوهش در تقاطع دو حوزه پردازش زبان طبیعی و بینایی ماشین، مسئله‌ی پرسش و پاسخ تصویری امروزه یکی از مسائل محبوب در این حوزه برای پژوهشگران است و کاربردهای فراوانی نیز دارد.

در سامانه پرسش و پاسخ تصویری، ورودی‌ها یک سوال مرتبط با تصویر و یک تصویر است که از این سیستم انتظار می‌رود که قادر باشد پاسخ صحیح سوال ورودی را پیش‌بینی کند. در این سامانه‌ها به دلیل اینکه میزان دقت پیش‌بینی پاسخ صحیح بسیار اهمیت دارد، پژوهشگران در تلاش هستند که با استفاده از شبکه‌های عصبی مناسب و سازوکارهایی مانند سازوکار توجه و ادغام مناسب ویژگی‌های استخراج شده ورودی‌ها، باعث بهبود میزان دقت پاسخ پیش‌بینی شده شوند [۲].

آموزش شبکه‌های عصبی مورد استفاده در سامانه پرسش و پاسخ تصویری دارای اهمیت فراوانی به جهت پیش‌بینی صحیح پاسخ‌ها است. هر چه مجموعه‌داده[5] مورد استفاده جامع‌تر و کامل‌تر باشد، آموزش مدل بهتر انجام شده و مدل کاراتر خواهد شد. به همین دلیل انتخاب و استفاده از مجموعه‌داده مناسب توسط پژوهشگران در این حوزه امری حیاتی محسوب می‌شود.

در سامانه‌های پرسش و پاسخ بصری از شبکه‌های عصبی گوناگونی می‌توان استفاده کرد. هم در زمینه‌ی پردازش تصویر و هم در زمینه‌ی پردازش متن، شبکه‌های عصبی وجود دارند که می‌توانند در کمک به این موضوع که خروجی این سامانه‌ها دقت خوبی داشته باشند، مفید واقع شوند و کارایی آن‌ها در پردازش ورودی‌ها است، به همین منظور در این سامانه‌ها از آن‌ها به صورت گسترده استفاده شده است. به طور کلی در بیشتر پژوهش‌ها و تحقیقات صورت گرفته در حوزه‌ی سامانه‌های پرسش و پاسخ تصویری، از شبکه‌های عصبی کانولوشنی با معماری‌های مختلف برای پردازش تصویر و از انواع مختلف شبکه‌های عصبی بازگشتی[6] جهت پردازش متن ورودی استفاده می‌شود.

ویژگی‌های موجود در شبکه‌ی عصبی کانولوشنی موجب شده است که این شبکه برای پردازش تصویر بسیار مناسب باشد و همچنین ساختار موجود در شبکه‌های عصبی بازگشتی این امکان را فراهم می‌کند که داده‌های متنی به خوبی در این شبکه‌ها پردازش شوند. در سامانه‌های پرسش و پاسخ تصویری باید داده‌های متنی به صورت بردارهای عددی به شبکه‌ی عصبی بازگشتی مورد استفاده داده

[5] Dataset
[6] Recurrent neural network(RNN)

شوند، تا مورد پردازش قرار بگیرند. به همین دلیل برای تبدیل لغات به بردارهای عددی در این سامانه‌ها باید از مدل‌های زبانی مانند Glove [8] و دیگر مدل‌ها در این زمینه استفاده کرد.

۳- کارهای مرتبط

برای حل مسئله پرسش و پاسخ تصویری و پیش بینی پاسخ مناسب برای سوال بر اساس تصویر ورودی، در سال‌های اخیر پژوهش‌های فراوانی انجام شده و رویکردهای مختلفی از جمله: رویکردهای مبتنی بر تعبیه‌گذاری مشترک[۷]، رویکردهای مبتنی بر سازوکار توجه و رویکردهای مبتنی بر مدل‌های ترکیبی معرفی شده‌اند.

در پژوهش Ma و همکاران [۳]، از سه شبکه‌ی عصبی کانولوشنی برای سیستم پرسش و پاسخ تصویری پیشنهادی خود، استفاده کرده‌اند. در مدل پیشنهادی پرسش و پاسخ تصویری در مقاله [۳] یک شبکه عصبی کانولوشنی برای پردازش متن ورودی، شبکه عصبی کانولوشنی دیگری برای پردازش تصویر ورودی استفاده شده و شبکه عصبی کانولوشنی آخر، یک شبکه عصبی کانولنشی چند حالته[۸] است که برای ادغام بازنمایی بدست آمده ورودی‌ها مورد استفاده قرار گرفته و بازنمایی مشترک را برای ورودی‌های سیستم پرسش و پاسخ تصویری تولید کرده و با استفاده از آن بازنمایی و تابع بیشینه هموار[۹] موجود در مدل، پاسخ نهایی در مدل پیش‌بینی می‌شود. برای ارزیابی مدل پیشنهادی مقاله [۳]، از دو مجموعه داده COCO-QA [9] و DAQUAR [۹] استفاده شده است که دقت صحت پاسخ پیش‌بینی شده در آن‌ها به ترتیب ۴۰/۵۴ درصد و ۷۶/۴۲ درصد است.

آندریاس و همکاران [۵]، برای پیش‌بینی پاسخ سوالات در مورد تصاویر، یک سیستم پرسش و پاسخ تصویری مبتنی بر ماژول‌های شبکه عصبی پیشنهاد کردند. در این سیستم ابتدا متن ورودی تعبیه‌سازی شده و سپس بردارهای عددی تولید شده به عنوان ورودی به شبکه‌ی حافظه‌ی کوتاه مدت طولانی[۱۰] جهت پردازش داده می‌شوند و ویژگی‌های سطح معنایی پرسش بدست می‌آید. ورودی تصویری نیز به یک شبکه عصبی کانولوشنی با معماری VGGNet [۱۰] از نوع ۱۶ لایه داده شده و ویژگی سطوح مختلف تصویر استخراج شده و بازنمایی مربوطه تولید و به ماژول‌ها داده می‌شود. در این مدل از تجزیه‌کننده متن[۱۱] استفاده شده و وظیفه این تجزیه‌کننده آنالیز متن تعبیه شده است که به عنوان ورودی دریافت می‌کند و نتیجه‌ی آن، انتخاب ماژول مناسب جهت پیش‌بینی پاسخ است. ماژول‌های موجود در این مدل مبتنی بر سازوکار توجه هستند. خروجی بدست آمده در این ماژول‌ها در انتهای مدل با بازنمایی متن ترکیب می‌شود و خروجی این ترکیب با استفاده از تابع بیشینه هموار پاسخ را با دقت ۵۵/۱ درصد پیش‌بینی می‌کند. ارزیابی مدل پیشنهادی بر روی مجموعه‌داده VQA1.0 [۹] انجام شده است.

یانگ و همکاران [۶]، برای پرسش و پاسخ تصویری یک مدل به نام SAN پیشنهاد کردند. این مدل بر مبنای دو سازوکار توجه پشت سر هم است. در این مدل تصویر توسط یک شبکه عصبی کانولوشنی با معماری VGGNet [۱۰] پردازش شده و ویژگی‌های آن

[7] Joint embedding
[8] Multimodal CNN
[9] Softmax function
[10] Long short term memory network (LSTM)
[11] Semantic parser

بدست می‌آید. پرسش ورودی بعد از تعبیه‌سازی به یک شبکه عصبی بازگشتی از نوع حافظه کوتاه مدت طولانی و یک شبکه عصبی کانولوشنی جهت پردازش داده می‌شود. بعد از بدست آمدن بازنمایی پرسش، از دو سازوکار توجه پشت سر هم استفاده می‌شود و نقاطی از تصویر که مرتبط با سوال هستند، پیدا و مشخص می‌شوند و طی چندین مرحله استدلال پاسخ مناسب در این مدل پیش‌بینی می‌شود.

کاظمی و همکاران [۷]، یک مدل بر پایه شبکه‌های عصبی و سازوکار توجه نرم[۱۲] برای پرسش و پاسخ تصویری پیشنهاد کردند. در این مدل در ابتدا تصویر به یک شبکه عصبی از نوع کانولوشنی با معماری ResNet [۱۰] با ۱۵۲ لایه داده شده و پردازش بر روی تصویر ورودی انجام می‌شود و ویژگی‌های آن استخراج می‌شود. همزمان با تصویر کلمات موجود در پرسش ورودی تعبیه شده و به یک شبکه عصبی برگشتی از نوع حافظه کوتاه مدت طولانی داده می‌شود. ویژگی‌های معنایی برای سوال بدست آمده و در مرحله بعد با ادغام ویژگی‌های بدست آمده برای ورودی‌ها، بازنمایی مشترک بدست آمده و به عنوان ورودی به دو لایه کانولوشن داده می‌شود. خروجی بدست آمده به یک تابع بیشینه هموار داده شده و با استفاده از این تابع وزن هر ویژگی سوال و تصویر مشخص می‌شود و بر روی تصویر توزیع توجه چندگانه انجام می‌شود. خروجی در این مرحله با بازنمایی حاصل شده از ورودی‌ها ادغام شده و بازنمایی مشترک تولید می‌شود و به عنوان ورودی به دو لایه تماما متصل وارد شده و با استفاده از تابع بیشینه هموار جواب پیش‌بینی می‌شود.

هاشمی و اصغری [۱]، یک مجموعه‌داده جدید برای پرسش و پاسخ تصویری به زبان فارسی معرفی کرده‌اند. آن‌ها از این مجموعه‌داده بر روی چندین مدل‌های مختلف پرسش و پاسخ تصویری استفاده کردند. برای بدست آورن مجموعه‌داده فارسی از مجموعه‌داده VQA1.0 استفاده شده و این مجموعه‌داده با مترجم گوگل و ترجمان ترجمه شده و در مدل‌ها مورد نظر، استفاده شده است.

در مدل اول پیاده‌سازی شده در مقاله [۱]، که یکی از ساده‌ترین مدل‌ها در حوزه پرسش و پاسخ تصویری است، ابتدا متن ورودی به یک شبکه عصبی کانولوشنی از نوع VGG-Net [۱۰] ۱۹ لایه داده می‌شود. سپس متن وردی جهت پردازش به یک شبکه عصبی برگشت‌پذیر حافظه کوتاه مدت طولانی داده می‌شود. در مرحله بعد خروجی‌های بدست آمده از دو شبکه عصبی استفاده شده به دو لایه تماما متصل داده می‌شوند و با کمک خروجی این دو لایه، بعد از عبور از یک لایه تماما متصل دیگر با استفاده از تابع بیشینه هموار، پاسخ پیش‌بینی می‌شود. با استفاده از مجموعه‌داده فارسی تولید شده در این مدل، زمانی که ترجمه مجموعه‌داده توسط ترگمان انجام می‌شود، دقت درستی پاسخ پیش‌بینی شده ۵۱/۳ درصد و زمانی که ترجمه مجموعه‌داده توسط گوگل انجام می‌شود، دقت درستی پاسخ پیش‌بینی شده ۵۰/۹۱ درصد خواهد بود.

مدل بعدی که در مقاله [۱] با استفاده از مجموعه‌داده فارسی پیاده‌سازی شده است، ابتدا تصویر ورودی را توسط یک شبکه عصبی کانولوشنی از نوع VGGNet [۱۰] ۱۶ لایه پردازش می‌کند. سپس از دو شبکه عصبی کانولوشنی و حافظه کوتاه مدت طولانی جهت پردازش متن ورودی استفاده می‌کند. در این مدل از دو سازوکار توجه استفاده شده و توسط این سازوکارها نقاط مرتبط با متن در تصویر مشخص شده و طی چندین مرحله استدلال پاسخ نهایی در این مدل بدست می‌آید. با استفاده از مجموعه‌داده فارسی تولید شده در این مدل، زمانی که ترجمه مجموعه‌داده توسط ترگمان انجام می‌شود، دقت درستی پاسخ پیش‌بینی شده ۵۲/۷۶ درصد و زمانی که ترجمه مجموعه‌داده توسط گوگل انجام می‌شود، دقت درستی پاسخ پیش‌بینی شده ۵۱/۳۸ درصد خواهد بود.

[12] Soft Attention Mechanism

مدل بعدی که در مقاله [۱] با استفاده از مجموعه‌داده فارسی پیاده‌سازی شده است، HieCo Attention نام دارد. در این مدل با مفهوم بازنمایی سلسله مراتبی سوال روبرو هستیم. در این روش بازنمایی در سه سطح کلمه، عبارت و کل متن اتفاق می‌افتد. همچنین در این مدل از یک سازوکار توجه به نام سازوکار CoAttention استفاده شده است که شباهت بین سوال و تصویر ورودی را محاسبه کرده و از آن در پیش‌بینی پاسخ مناسب استفاده می‌کند. با استفاده از مجموعه‌داده فارسی تولید شده در این مدل زمانی که ترجمه مجموعه‌داده توسط ترگمان انجام می‌شود، دقت درستی پاسخ پیش‌بینی شده ۵۱/۸۵ درصد و زمانی که ترجمه مجموعه‌داده توسط گوگل انجام می‌شود، دقت درستی پاسخ پیش‌بینی شده ۴۸/۰۷ درصد خواهد بود.

۴-مجموعه‌داده

همانگونه که می‌دانید برای پرسش و پاسخ تصویری مجموعه‌داده عمومی به زبان فارسی وجود ندارد. برای طراحی یک سیستم پرسش و پاسخ تصویری به زبان فارسی، نیاز است یک مجموعه‌داده به زبان فارسی به وجود آورد و از آن استفاده کرد. برای همین منظور ما از مجموعه‌داده پرکاربرد VQA.1 [۹] استفاده کردیم که به زبان انگلیسی است. در مرحله اول، این مجموعه‌داده را توسط مترجم گوگل ترجمه کرده و سؤالات و پاسخ‌های موجود در آن را به زبان فارسی تبدیل کردیم تا مجموعه‌داده ما به زبان فارسی شود، سپس داده‌ها را هم زده و ۸۰ درصد داده‌ها را برای آموزش و ۲۰ درصد داده‌ها را برای اعتبارسنجی در نظر گرفتیم و فرآیند آموزش مدل با مجموعه‌داده فارسی بوجود آمده را انجام دادیم.

۴-۱ پیش‌پردازش مجموعه‌داده

بعد از بدست آوردن مجموعه‌داده فارسی به وسیله مترجم و تقسیم آن به دو قسمت که در بخش قبل گفته شد، نوبت به پیش‌پردازش داده‌ها و آماده‌سازی آن‌ها به جهت پردازش در مدل پیشنهادی می‌رسد. در این بخش داده‌های تصویری تغییر اندازه داده می‌شوند و به سایز ۲۲۴×۲۲۴ تبدیل می‌شود. همچنین متن پرسش و کلمات موجود در آن Tokenize شده و طول سوالات، ۲۰ کلمه در نظر گرفته می‌شود.

۵- روش پیشنهادی

در این مقاله یک سیستم پرسش و پاسخ تصویری به زبان فارسی معرفی خواهد شد. روش پیشنهادی ما به سه مرحله تقسیم می‌شود. مرحله اول پردازش تصویر ورودی، مرحله دوم پردازش سوال ورودی و مرحله سوم ادغام بازنمایی بدست آمده برای ورودی‌ها و استفاده از لایه‌های مختلف جهت پیش‌بینی پاسخ نهایی است که در ادامه به تشریح این مراحل خواهیم پرداخت.

۵-۱ مرحله اول

- پردازش تصویر ورودی

در این مرحله پس از پیش‌پردازش تصویر ورودی، آن را به یک شبکه عصبی کانولوشنی با معماری ResNext [۱۱] با ۱۰۱ لایه می‌دهیم. تصویر ورودی در این شبکه عصبی پردازش می‌شود و ویژگی‌های تصویر در ابعاد ۲۰۴۸×۷×۷ استخراج می‌شوند. سپس برای اینکه اندازه ویژگی‌های سؤال و تصویر یکسان شوند، ابعاد ویژگی‌های تصویر را به صورت ۵۱۲×۷×۷ تغییر می‌دهیم. با مطالعات فراوان

بر روی معماری‌های مختلف شبکه عصبی کانولوشنی، به دلیل اینکه تا به امروز یکی از بهترین معماری‌های این شبکه برای تشخیص اشیاء، معماری ResNext [۱۱] است و یکی از ویژگی‌های آن دقت بالا با وجود پارامترهای کم است، این معماری با ۱۰۱ لایه را برای پردازش تصویر در مدل پیشنهادی پرسش و پاسخ تصویری خود انتخاب کردیم.

۵-۲ مرحله دوم

- پردازش پرسش ورودی

در این مرحله پس از پیش‌پردازش پرسش ورودی، کلمات موجود در پرسش به یک لایه تعبیه‌سازی به عنوان ورودی داده می‌شوند. کلمات متن در این لایه، تعبیه‌سازی شده و به بردارهای عددی تبدیل می‌شوند. این بردارهای عددی به عنوان ورودی به یک سازوکار توجه‌به‌خود [۱۲] داده می‌شوند و توسط این سازوکار، ارتباط بین لغات در متن ورودی جهت درک و فهم بهتر پرسش ورودی بدست می‌آید و خروجی این سازوکار به عنوان ورودی به یک شبکه عصبی بازگشتی از نوع حافظه کوتاه مدت طولانی داده شده و در این شبکه پرسش ورودی پردازش می‌شود و ویژگی‌های آن بدست می‌آید.

- استفاده از سازوکارهای توجه

در مدل پیشنهادی از سازوکارهای توجه به منظور افزایش دقت درستی پاسخ پیش‌بینی شده، استفاده شده است. سازوکارهای مورد استفاده در مدل پیشنهادی این مقاله، سازوکار توجه‌به‌خود و سازوکار توجه‌چندسر[۱۳] [۱۳] است.

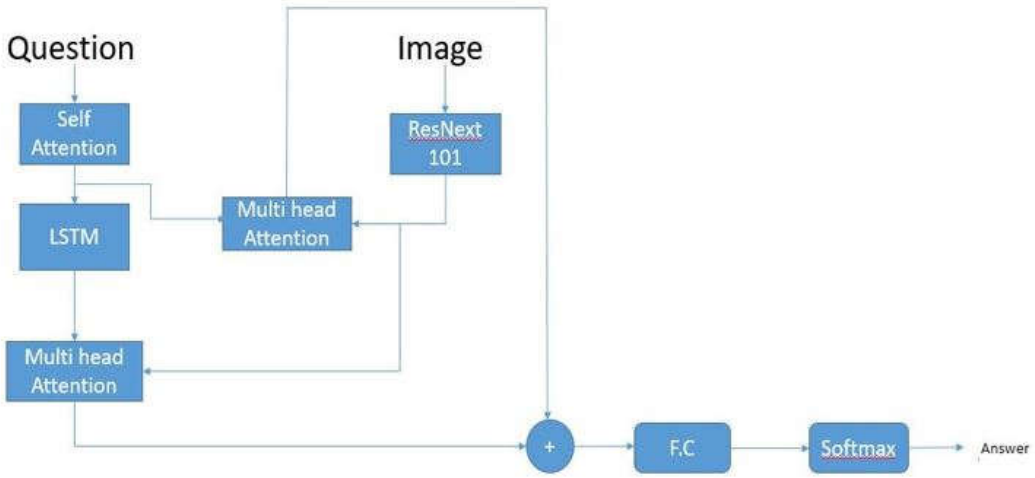
۵-۳ مرحله سوم

- ترکیب ویژگی‌ها و پیش‌بینی پاسخ نهایی

برای ترکیب کردن ویژگی‌های تصویر با ویژگی‌های پرسش و کلمات، از سازوکار توجه چندسر استفاده می‌شود. به این صورت که ابتدا ویژگی‌های تصویر و ویژگی‌های کلمات (خروجی سازوکار توجه‌به‌خود) در توجه‌چندسر شرکت می‌کنند. این کار باعث می‌شود کلمات با توجه به ویژگی‌های تصویر پردازش شوند.

همچنین به طور موازی، ویژگی‌های استخراج شده پرسش توسط حافظه کوتاه مدت طولانی، به همراه ویژگی‌های استخراج شده تصویر در سازوکار توجه‌چندسر دیگری پردازش می‌شوند. در نهایت خروجی سازوکار توجه‌چندسر کلمات-تصویر و پرسش-تصویر با یکدیگر جمع شده و با استفاده از نرمال سازی L2، لایه تماما متصل و تابع بیشینه هموار منجر به پیش‌بینی خروجی می‌شوند. در شکل ۲، ساختار مدل پیشنهادی در این مقاله برای پرسش و پاسخ تصویری نشان داده شده است.

[13] Multihead attention



شکل ۲: ساختار مدل پیشنهادی پرسش و پاسخ تصویری

۶- جزئیات پیاده‌سازی

در این بخش به بیان جزئیات پیاده‌سازی مدل پیشنهادی خواهیم پرداخت. این جزئیات به صورت زیر است:

- در پیاده‌سازی مدل، از الگوریتم بهینه‌سازی ADAM [۱۴] استفاده شده است.
- استفاده از تکنیک دوریزی[۱۴] [۱۵] با نرخ ۰.۳
- استفاده از نرمال‌سازی دسته‌ای[۱۵] [۱۶]
- نرخ یادگیری[۱۶] ۰.۰۰۰۱ است.
- در این پژوهش شبکه در پنجاه دوره آموزش می‌بینند.

۶-۱ محیط پیاده‌سازی

انتخاب محیط مناسب و همچنین استفاده از کتابخانه‌های مناسب برای پیاده‌سازی شبکه‌های عصبی امری مهم و پراهمیت است. یکی از زبان‌های برنامه‌نویسی بسیار کارا در این حوزه، زبان برنامه‌نویسی پایتون[۱۷] است که پیاده‌سازی مدل پیشنهادی در این مقاله توسط

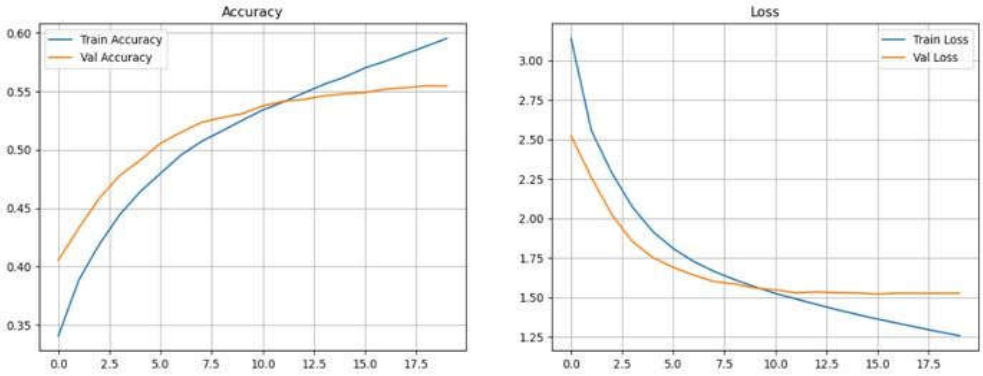
[14] Drop out
[15] Batch normalization
[16] Learning rate
[17] Python

این زبان انجام می‌شود. همچنین از دو کتابخانه معروف و در عین حال قدرتمند به نام‌های TensorFlow [۱۷] و Keras [۱۸] در پیاده‌سازی مدل پیشنهادی استفاده کردیم. دلیل انتخاب این دو کتابخانه کارایی بالای آن‌ها در حوزه کاری این پژوهش و در کل در حوزه یادگیری عمیق است. این دو کتابخانه موجب افزایش سرعت می‌شوند و امکان استفاده از حداکثر توان GPU را برای ما فراهم کرده در نتیجه زمانی که مجموعه‌داده مورد استفاده بزرگ باشد، این ویژگی بسیار مفید خواهد بود [۱۸].

۷- نتایج بدست آمده مدل پیشنهادی

در مدل پیشنهادی ما، سه دسته پرسش مد نظر است. در ابتدا پرسش‌هایی که جواب بله و خیر دارند. دسته دوم پرسش‌هایی که جواب عددی دارند و دسته سوم دیگر پرسش‌ها (پرسش‌هایی که جواب تک‌کلمه‌ای یا دوکلمه‌ای دارند) هستند. نتایج حاصل شده در این سه دسته پرسش در این بخش بیان می‌شود و در نهایت دقت کل مدل بیان خواهد شد.

برای ارزیابی مدل پیشنهادی این پژوهش از مجموعه‌داده ساخته شده فارسی استفاده خواهد شد. همانگونه که گفته شد، این مجموعه‌داده با استفاده از مجموعه‌داده VQA1.0 [۹] و مترجم گوگل ساخته شده است. در جدول ۱، نتایج حاصل شده برای مدل پیشنهادی با استفاده از مجموعه‌داده فارسی نشان داده شده است. همچنین در شکل ۳، نمودار نمودار دقت و Loss مدل پیشنهادی نشان داده شده است.



شکل ۳: نمودار دقت و Loss مدل پیشنهادی

جدول ۱: نتایج مدل پیشنهادی

نوع پاسخ	بله/خیر	اعداد	غیره	دقت کل
دقت پیش بینی پاسخ برحسب درصد	۷۷/۸۵	۳۹/۱	۴۶/۲	۵۸/۶

۸- ارزیابی مدل پیشنهادی

تنها پژوهش موجود در زمینه پرسش و پاسخ تصویری به زبان فارسی، مقاله [۱] است که از مجموعه‌داده‌ای مشابه با مجموعه‌داده مورد استفاده در این پژوهش، بر روی مدل‌ها و معماری‌های مختلف پرسش و پاسخ تصویری استفاده کرده است. ما یک معماری جدید برای سیستم پرسش و پاسخ تصویری به زبان فارسی پیشنهاد کرده‌ایم و در این بخش نتایج بدست آمده برای مدل پیشنهادی خود را با نتایج بدست آمده در دیگر مدل‌های پرسش و پاسخ تصویری به زبان فارسی [۱] مقایسه خواهیم کرد. در جدول ۲، مقایسه‌ای بین نتایج حاصل در مدل پیشنهادی این مقاله با نتایج حاصل شده از مدل‌های دیگر پرسش و پاسخ تصویری به زبان فارسی [۱] آورده شده است (مجموعه‌داده مورد استفاده در تمامی مدل‌ها یکسان بوده و ترجمه مجموعه‌داده VQA1.0 [۹] به فارسی با استفاده از گوگل است).

جدول۲: نتایج مدل پیشنهادی در این مقاله و دیگر مدل‌های پرسش و پاسخ تصویری به زبان فارسی

	نوع پاسخ			
روش	بله/ خیر	اعداد	غیره	دقت کل
Lstm Q + VGG 19 [۱]	۷۶/۱۴	۳۲/۹۷	۳۵/۷۸	۵۰/۵۳
BiLSTM + RESNET 152 [۱]	۷۶/۴۶	۳۱/۶۳	۳۸/۶	۵۱/۸۹
Lstm Q + RESNET 152 [۱]	۷۶/۸۳	۳۱/۷۵	۳۸/۷۷	۵۲/۱۳
CNNQ + RESNET 152 [۱]	**۷۸/۳۴**	۳۱/۹۱	۳۸/۹۸	۵۲/۸۲
SAN_LSTM [۱]	۷۷/۸۳	۳۳/۱۹	۳۹/۰۸	۵۲/۸۴
SAN_CNN [۱]	۷۷/۴۹	۳۳/۱۷	۳۹/۱۸	۵۲/۷۶
CoAttention [۱]	۷۶/۶۲	۳۲/۷	۳۸/۱۲	۵۱/۸۵
مدل پیشنهادی	۷۷/۸۵	**۳۹/۱**	**۴۶/۲**	**۵۸/۶**

همانگونه که در جدول ۲، مشخص است، مدل پیشنهادی ما در دو دسته سوال، یعنی دسته‌ی سوالات مربوط به اعداد و سوالات غیره نسبت به تمام مدل‌ها پیاده‌سازی شده با مجموعه‌داده فارسی در مقاله [۱] به درصد دقت بالاتری دست یافته است و تنها در سوالاتی که پاسخ بله و خیر دارند، نسبت به یکی از مدل‌های پیاده‌سازی شده در مقاله [۱] درصد دقت کمتری را دارد. همچنین دقت کل در مدل پیشنهادی ما از تمام مدل‌های پیشنهادی در مقاله [۱] بیشتر بوده است و از آنجایی که مقاله [۱] تنها پژوهش انجام شده در حوزه پرسش و پاسخ تصویری به زبان فارسی است، پس مدل ما بهترین مدل موجود پرسش و پاسخ تصویری به زبان فارسی است.

۹- نتیجه‌گیری

در این مقاله یک سیستم پرسش و پاسخ تصویری به زبان فارسی ارائه شده است. در این سیستم، تصویر ورودی به یک شبکه عصبی کانولوشنی از نوع ResNext ۱۰۱ [۱۱] لایه داده شد و ویژگی‌های تصویر استخراج شد و ورودی متنی یعنی پرسش، در ابتدا بعد از تعبیه‌سازی به سازوکار توجه‌به‌خود داده شد و خروجی این سازوکار به یک شبکه عصبی بازگشتی از نوع حافظه کوتاه مدت طولانی

داده شده و پرسش ورودی نیز پردازش و ویژگی‌های آن استخراج شده و بازنمایی پرسش بدست آمد. سپس بازنمایی بدست آمده برای پرسش و تصویر ورودی به یک سازوکار توجه‌چندسر داده می‌شود، همچنین بازنمایی تصویر با خروجی سازوکار توجه‌به‌خود به یک سازوکار توجه چند سر دیگر داده می‌شود. خروجی هر دو سازوکار توجه‌چندسر با یکدیگر جمع شده و بعد از عبور از لایه تماما متصل و با استفاده از تابع بیشینه هموار جواب خروجی پیش‌بینی می‌شود. ارزیابی مدل پیشنهادی در این مقاله با استفاده از مجموعه‌داده فارسی انجام می‌شود. این مجموعه‌داده توسط ما با استفاده از مجموعه‌داده VQA [۹] (نسخه اول) و مترجم گوگل بدست آمد. در نهایت نتایج بدست آمده نشان دهنده‌ی این موضوع است که مدل ما بهترین نتیجه یعنی بیشترین دقت درستی پاسخ پیش‌بینی شده در سیستم پرسش و پاسخ تصویری فارسی را بدست آورده است. برای بهبود دقت در سیستم پرسش و پاسخ تصویری به زبان فارسی، پژوهشگران این حوزه می‌توانند از معماری‌های دیگر شبکه عصبی کانولوشنی جهت پردازش تصویر استفاده کنند. همچنین استفاده از ترنسفرمرها و ساخت مجموعه‌داده‌های بزرگتر به زبان فارسی، می‌تواند باعث بهبود دقت در سیستم پرسش و پاسخ تصویری به زبان فارسی شود.

مراجع

[۱] هاشمی، اصغری؛ مریم السادات، علیرضا، " پرسش و پاسخ تصویری در فارسی"، دانشگاه علم و صنعت ایران، مجله یادگیری عمیق، 1399.

[2] 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.

[3] L. Ma, Z. Lu, and H. Li. "Learning to Answer Questions From Image using Convolutional Neural Network". In Proc. Conf. AAAI, 2016.

[4] M. Ren, R. Kiros, and R. Zemel, "Exploring models and data for image question answering", in Advances in neural information processing systems, 2015.

[5] J. Andreas, M. Rohrbach, T. Darrell, and D. Klein. "Deep Compositional Question Answering with Neural Module Networks" . In Proc. IEEE Conf. Comp. Vis. Patt. Recogn., 2016.

[6] Z. Yang, X. He, J. Gao, L. Deng, and A. Smola. "Stacked Attention Networks for Image Question Answering". In Proc. IEEE Conf. Comp. Vis. Patt. Recogn., 2016.

[7] V. Kazemi, A. Elqursh. "Show, Ask, Attend, and Answer: A Strong Baseline For Visual Question Answering." In: arXiv preprint arXiv: 1704.03162v2, 2017.

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

[9] M. Malinowski and M. Fritz. "A multi-world approach to question answering about real-world scenes based on uncertain input." In Proc. Advances in Neural Inf. Process. Syst., pages 1682–1690, 2014.

[10] A.Krizhevsky, I.Sutskever, G.Hinton. "ImageNet classification with deep convolutional neural networks" Communications of the ACM. 60 (6): 84–90.2017.

[11] S.xie, R.Gitshick, P.Dollar, Z.Tu and K.He. "Aggregated Residual Transformations for Deep Neural Networks". EEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.

[12] A. Ambartsoumian, F. Popowich. "Self-Attention: A Better Building Block for Sentiment Analysis Neural Network Classifiers", Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, 2018.

[13] Z.Niu, G.Zhong, H.Yiu. " A review on the attention mechanism of deep learning", Neurocomputing, Volume 452, Pages 48-62, 2021.

[14] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," International Conference on Learning Representations, 2014.

[15] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," Journal of Machine Learning Research, vol. 15, pp. 1929–1958, (2014).

[16] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," CoRR, vol. abs/1502.03167, 2015.

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

[18] R.Conlin, K.Erickson, J.Abbate, E.Koleman. "Keras2c: A library for converting Keras neural networks to real-time compatible C". Engineering Applications of Artificial Intelligence 100: 21–40. 2020.

Visual Question Answering in Persian Based on deep learning techniques

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

Abstract

These days, image question and answer systems are widely used in order to automatically answer questions related to the content of images. It is possible to use a video question and answer system to predict the correct answer to a question based on an image and a text question about the image as input. Ideally, these systems should predict answers with high accuracy. Various factors, including the choice of appropriate neural networks and the choice of appropriate datasets, play an important role in achieving this goal. Additionally, different attention mechanisms can be used in the model to improve its performance. Few studies have been conducted on visual question and answer systems in Persian. Therefore, we introduce a visual question and answer system in Persian in this article. We used convolutional neural networks with ResNext architecture for image processing in the proposed model, which was used for the first time in video question and answer applications. We also used a recurrent neural network of the type of long-term and bilateral short-term memory to process the input text. As part of the proposed model, two types of attention mechanisms are employed. The results of this study demonstrate that the predicted answer in the proposed model of this article is the most accurate among the Persian examples

Visit this paper code at: https://github.com/amirshnll/persian-visual-question-answering

Keywords: visual question and answer system, convolutional neural network, recurrent neural network, attention mechanism