مجله علمی مطالعات پژوهشی در مهندسی برق آینده، ۱۴۰۳، دوره ۱، شماره ۱، ص ص ۱-۱۲

مجله علمی مطالعات پژوهشی
در مهندسی برق آینده
www.journalhi.com/pow/

بهبود پاسخ دقیق در مورد اقدامات انسانی در سیستم پرسش و پاسخ تصویری

امیر ساجدی [۱]

۱- دکتری مهندسی برق، دانشگاه پیام نور، شیراز، ایران.

اطلاعات مقاله

مقاله پژوهشی کامل
دریافت: ۲۷ اسفند ۱۴۰۱
پذیرش: ۱۰ اردیبهشت ۱۴۰۲
ارائه در سایت: ۱۲ خرداد ۱۴۰۲

کلید واژگان:
پردازش تصویر
شناسایی فعالیت انسان
یادگیری عمیق
شبکه عصبی کانولوشنی
سیستم پرسش دیداری پاسخ
حافظه بلند مدت

چکیده

تشخیص فعالیت یک ابزار ضروری است که تحلیلی کارآمد روی رفتار انسان و نیز تعاملات کارآمد بین انسان‌ها و سیستم‌های بینایی را ممکن می‌سازد. درعین‌حال شناخت عمل انسانی به دلیل تغییرات قابل توجه در اقدامات انسانی، از لحاظ سبک‌های شخصی، ظاهر انسان، دیدگاه دوربین، پس‌زمینه متغیر و دیگر تغییرات محیطی یک وظیفه بسیار چالش‌برانگیز است، اما تاکنون یک مدل موثر و کارآمد که بتواند با محاسبات کم مراحل تشخیص و درک تصویر را آشکار کند، ارائه نشده‌است. این حوزه از تحقیق به دلیل کاربرد در حوزه‌های مختلف اعم از پزشکی، تعامل انسان-رایانه، تجاری و...، نظر بسیاری از محققان را به سمت خود جلب کرده‌است. برای این منظور، ما با هدف بهبود دقت پاسخ در سیستم پرسش و پاسخ تصویری[۱] و افزایش میزان دقت در تشخیص فعالیت انسان، راه‌کار استفاده از مکانیسم یادگیری عمیق به جهت تشخیص اطلاعات تصاویر و سیستم پرسش و پاسخ تصویری به جهت پیش‌بینی پاسخ از سوالات موجود در تصاویر را پیشنهاد می‌کنیم. روش پیشنهادی ما چارچوب شناسایی حلقه‌ها را مقیاس گذاری، استقرار و نظارت می‌کند. در نهایت پاسخ به سوال با توجه به وزن‌های محاسبه شده در مرحله آموزش، ویژگی‌های تصویر ورودی و مبنای درستی محاسبه می‌شود و نهایتا با توجه به خوشه‌بندی پیشنهادی موجب بهبود در پاسخ دقیق می‌شود. روش پیشنهادی در دقت و سرعت اجرای الگوریتم نسبت به سایر روش‌ها برتری دارد.

Improving the accurate answer about human actions in the visual question and answer system

Amir Sajedi[1]

1- PhD in Electrical Engineering, Payam Noor University, Shiraz, Iran.

Article Information

Original Research Paper
Received 18 March 2023
Accepted 02 October 2023
Available Online 04 October 2023

Keywords:
Image Processing
Identification of human activity
deep learning
Convolutional neural network
Visual question and answer system
long-term memory

Abstract

Activity detection is a necessary tool that enables efficient analysis of human behavior as well as efficient interactions between humans and vision systems. At the same time, recognition of human action is a very challenging task due to significant changes in human actions, in terms of personal styles, human appearance, camera view, changing background and other environmental changes, but so far an effective and efficient model that can recognize with few steps calculations and reveal the understanding of the image, is not presented. This field of research has attracted the attention of many researchers due to its application in various fields such as medicine, human-computer interaction, business, etc. For this purpose, with the aim of improving the accuracy of answers in the image question and answer system and increasing the accuracy of human activity detection, we have developed a solution of using deep learning mechanism to recognize the information of images and the image question and answer system to predict the answers to the questions in We suggest pictures. Our proposed method scales, deploys and monitors the ring detection framework. Finally, the answer to the question is calculated according to the weights calculated in the training stage, the features of the input image and the basis of correctness, and finally, according to the proposed clustering, it leads to an improvement in the accurate answer. The proposed method is superior in accuracy and speed of algorithm execution compared to other methods.

[1] Visual Question Answering

۱- مقدمه

از سال ۱۹۹۱ پردازش و تحلیل فعالیت انسان، یکی از مهمترین موضوعات بینایی ماشین بوده‌است. این امر سبب ارائه سیستم‌های نظارتی در زمینه‌های مختلف و مشاغل گوناگون، بیمارستان‌ها و غیره شده‌است. پردازش فعالیت‌ها شامل شناسایی الگوهای حرکتی و تولید توصیفات سطح بالا از این دسته از فعالیت‌ها می‌شود. پژوهش‌های بسیاری برای شناسایی فعالیت انسان انجام شده‌است. در این زمینه، تکنیک‌های ردیابی و شناسایی حرکات انسان و اندام‌های مختلف بدن به کار برده می‌شوند. تمامی‌روش‌ها بر اساس ردیابی حرکات اجزای مختلف بدن انسان مانند دست‌ها می‌باشند و این حرکات از طریق دوربین‌های مختلف دنبال می‌شوند. همچنین در برخی از پژوهش‌ها شناسایی فعالیت‌ها از مدل‌های سه بعدی و مدل‌های دو بعدی استخراج می‌شوند. مدل‌های پویا در تشخیص، فعالیت‌ها را به فعالیت‌های ساده و پیچیده دسته‌بندی می‌کنند [1]. در تمامی پژوهش‌ها اجزای مهم بدن انسان در بینایی ماشین مانند دست، پا و سر دنبال می‌شوند. حرکات آنها تشخیص داده شده و با ردیابی آن‌ها نوع فعالیت توصیف می‌گردد. در واقع تشخیص فعالیت‌های انسانی یک ابزار ضروری است که تحلیلی کارآمد روی رفتار انسان و نیز تعاملات کارآمد بین انسان‌ها و سیستم‌های بینایی را ممکن می‌سازد. در عین‌حال شناخت عمل انسانی به دلیل تغییرات قابل توجه در اقدامات انسانی، از لحاظ سبک‌های شخصی، ظاهر انسان، دیدگاه دوربین، پس‌زمینه متغیر و دیگر تغییرات محیطی یک وظیفه بسیار چالش‌برانگیز است. هنوز هم به رسمیت شناختن عملی در تصاویر، یک حوزه تحقیق فعال در بینایی ماشین و تشخیص الگوها است. با توجه به‌اهداف و نیازهای بشری، اینترنت مجموعه عظیمی‌از تصاویر را شامل می‌شود و بسیار هیجان‌انگیز است که اعمال انسانی را در تصاویر آن آنالیز کنیم [2]. در ادراک بصری انسان، به رسمیت شناختن یک عمل انسانی تنها با تمرکز بر بخش‌های کلیدی خاص بدن انسان و یا هدف عمل انجام می‌شود [3]. با دسته‌بندی رفتار می‌توان نیازهای مختلف از جمله عکس‌العمل مناسب را شناسایی کرده و اقداماتی سریعتر در راستای رفع نیازها انجام داد. در کل حتی اگر اجرای ما فقط ورودی سنسور را در یک زمان اجرای خاص تبلیغ کند و با توجه به این واقعیت که توالی‌ها ممکن است به دقت معقولی رسیده باشند. خصوصیات موقت در داده‌های حسگر قرار می‌گیرد، بنابراین اجازه می‌دهد مدل بدست‌آید. ما می‌دانیم که افراد غیرحرفه‌ای ممکن است در واقع پیش بینی رفتار انسان را فقط با استفاده از چند نمونه "آموزش" بهتر انجام دهند، در حالی که کارهای قبلی به طور معمول تفسیرپذیری یا نقش آن‌ها را در بهبود اعتماد به نفس ارزیابی کرده است. فرضیه اولیه ما این است که، کارهای زیادی برای توسعه روش‌های بهبود یافته برای بهبود توانایی تیم‌های AI-human انجام شده است. پرسش و پاسخ یکی از مسایل مهم در مبحث پردازش زبان طبیعی است، که با دو روش درک مطلب و انتخاب پاسخ مناسب به نتیجه می‌انجامد. در گذشته حل مسائل پردازش زبان طبیعی از جمله پرسش و پاسخ، براساس روش‌های آماری بوده و محققان مجموعه‌هایی از ویژگی‌ها را براساس متن ورودی، تولید می‌کردند[4] . یکی از رویکردهای پاسخ به سوالات بصری، پیش‌بینی پاسخ از سوالات مطرح شده‌است. در معیارهایی مانند تشخیص تصویر با استفاده از شبکه‌های عصبی کانولوشن عمیق، سیستم‌های تشخیص اشیاء دچار چالش‌های متعددی شده‌است و یک روش طبیعی برای حل این دسته‌از مسائل، ساخت سیستمی است که به آن تصویر و سوال مبتنی بر متن داده می‌شود و نهایتا در خروجی یک پاسخ مبتنی بر متن را ارائه می‌دهد[5] .با توجه به مطالعات گسترده روی نقاط مورد توجه انسان در انتخاب پاسخ به سوالات بصری و درک بخش مورد توجه در پاسخ به سوال پرسیده شده، نتایج ارزشمندی از آن حاصل می‌گردد[6] . با توجه به یک تصویر و یک سوال از زبان طبیعی در مورد تصویر، وظیفه اصلی سیستم ارائه یک پاسخ دقیق به زبان طبیعی است. منعکس کننده سناریوهای دنیای واقعی، کمک به‌افراد کم بینا است، هر دو پرسش و پاسخ بی پایان هستند. در نتیجه، سیستمی‌که در آزمون VQA موفق شود معمولاً به درک دقیق جزئیات تصویر و استدلال پیچیده نسبت به سیستمی‌که زیرنویس‌های تصویر عمومی را تولید می‌کند، نیاز دارد. به طور خاص، تحقیق در زیرنویس تصویر و فیلم که ترکیبی از بینش کامپیوتر، پردازش زبان طبیعی و بازنمایی و استدلال دانش است در گذشته بطور چشمگیری افزایش یافته‌است [7]. شبکه‌های عصبی عمیق با ترکیبی جدید از یادگیری کنترل شده از اقدامات انسانی و یادگیری تقویت شده از انواع سرگرمی‌ها آموزش داده می‌شوند. همچنین الگوریتم جستجوی جدیدی را معرفی می‌کنیم که شبیه سازی مونت کارلو را با شبکه‌ای از مقادیر و سیاست ترکیب می‌کند [8].

با استفاده از LSTM جداگانه برای هر قهرمان و بدون داده‌های اقدامات انسانی، استراتژی‌های قابل تشخیص را می‌آموزد. این نشان می‌دهد که یادگیری تقویتی[1] می‌تواند برنامه ریزی طولانی مدت با مقیاس بزرگ اما قابل دستیابی - بدون پیشرفت اساسی، بر خلاف انتظارات ما از شروع پروژه، انجام دهد[9] .

انسان‌ها سطوح درک بصری را نشان می‌دهند که فراتر از فرمول بندی‌های فعلی وظایف اصلی بینایی است. عنصر کلیدی هوش بصری توانایی تعامل با محیط و یادگیری از اینگونه تعاملات است. مدل‌های پیشرفته فعلی در بینایی رایانه با استفاده از تصاویر ثابت یا فیلم‌ها آموزش می‌بینند. این با یادگیری انسان متفاوت است [10].

پاسخ به سوال تجسم یافته (Embodied QA) - جایی که یک عامل در مکانی تصادفی در یک محیط سه بعدی گسترش می‌یابد و سوال ("ماشین چه رنگی است؟") را می‌پرسد. برای پاسخگویی، گزینه منتخب ابتدا باید با هوشمندی به جستجوی محیط بپردازد، اطلاعات دیداری لازم را از طریق دید خودمحوری جمع آوری نموده و سپس به سوال ("نارنجی") پاسخ دهد.مکانیزم Embodied QA به طیف وسیعی از مهارت‌های هوش مصنوعی نیاز دارد از جمله درک زبان، شناخت بصری، درک فعال، ناوبری با هدف، استدلال منطقی، حافظه طولانی مدت و زبان مبنی بر اقدامات انسانی [11]. مثال‌های روشنی برای آموزش بازی‌های آتاری جهت بازی کردن و با استفاده از داده‌های پیکسلایز خام و به دست آوردن مهارت‌های پیشرفته در استفاده از ورودی‌های حسی خام موجود است. از این رو، به دلیل کمبود یک معیار معمول پذیرفته شده، تعیین کمیت در حوزه کنترل مداوم دشوار بوده‌است [12]. یادگیری تقویت (RL) شاخه‌ای از یادگیری ماشین است که مربوط به تصمیم گیری توالی است. RL دارای یک نظریه ریاضی غنی بوده و انواع کاربردهای عملی را در بر دارد [13] [14]. آموزش شبکه‌های عصبی عمیق‌تر دشوارتر است. ما یک چارچوب یادگیری باقی‌مانده برای سهولت آموزش شبکه‌هایی که عمیق‌تر از شبکه‌های قبلی هستند، ارائه می‌دهیم. ما به جای یادگیری توابع بدون مرجع، لایه‌ها را به عنوان یادگیری توابع باقیمانده و با اشاره به ورودی‌های لایه به طور صریح اصلاح می‌کنیم. یک شبکه عصبی منفرد، بسته‌های محدود کننده و مبتنی بر احتمالات مربوط به کلاس را مستقیماً از تصاویر کامل در یک ارزیابی پیش‌بینی می‌کند. از آنجا که کل خط مسیر شناسایی یک شبکه واحد است، می‌توان آن را از انتها به انتهای

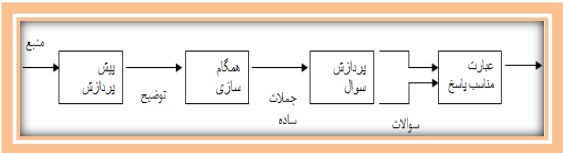
[1] Reinforcement Learning

دیگر و مستقیماً بر روی عملکرد تشخیص بهینه کرد [15] [16] [17] [18]. امروزه نیاز به تحلیل و بررسی فعالیت‌های انسان به موضوعی مهم تبدیل شده است. ایجاد بستری مناسب که با دقت بالا و هزینه پایین به آنالیز اعمال انسانی در تصاویر بپردازد می‌تواند موجب حذف برخی مشاغل و یا در عین حال بالا بردن میزان دقت در خروجی عمل گردد. تاکنون پژوهش‌های قابل‌توجهی با ارائه راه حل‌هایی متفاوت در تحلیل و بررسی این فعالیت‌ها صورت پذیرفته که شامل تجزیه و تحلیل رفتار، سطح عملکرد، پیش‌بینی عمل، پیشنهاد روش و بهبود عملکرد می‌باشد. در حقیقت، انواع مختلفی از تشخیص رفتار در سیستم‌های سنتی و همچنین پیشرفته در تشخیص و تحلیل رفتار در دسترس است. به طور مشابه با توسعه مدل‌های یادگیری ماشینی و اجرای موفقیت‌آمیز آن‌ها برای رفع مشکلات مهم حوزه‌های مختلف، تشخیص انواع رفتار از طریق یادگیری ماشین و رویکردهای مبتنی بر هوش مصنوعی نیز مورد توجه محققان قرار گرفته است. نیاز به یادگیری ماشین در این زمینه به این واقعیت منجر می‌شود که بررسی تفاوت افراد در اعمال یکسان، زمینه پیچیده‌ای را مورد بررسی قرار می‌دهد. در این تحقیق هدف ما بهبود دقت پاسخ در سیستم پرسش و پاسخ تصویری و افزایش میزان دقت در تشخیص فعالیت انسان است. دراین راستا از مکانیسم یادگیری عمیق به جهت تشخیص اطلاعات تصاویر و سیستم پرسش و پاسخ تصویری به جهت پیش‌بینی پاسخ از سوالات موجود در تصاویر خواهیم پرداخت.

مواد و روش کار

سهم عمده‌این پژوهش در معرفی دیتاست‌های لازم به‌این شرح است که: ابتدا یک مجموعه داده پاسخ سوال بصریVQA شامل بیش از ۵۰۰۰ تصویر و ۳۰۰۰۰ بسته دوتایی سوال-پاسخ ایجاد می‌شود. تا تحقیقات VQA تقویت شود. از نظر توانایی دانش، این اولین مجموعه داده برای VQA است. ما همچنین یک خط ارتباطی نیمه اتوماتیک ایجاد می‌کنیم تا بتواند به طور موثر مجموعه داده‌های VQA را از کتاب‌های آموزشی و کتابخانه‌های دیجیتال آنلاین ایجاد کند. خط ارتباطی ما می‌تواند به طور گسترده‌ای در سایر حوزه‌های تصویربرداری حتی فراتر از آسیب شناسی اعمال شود، مانند رادیولوژی، سونوگرافی و غیره.

ما چندین روش VQA کاملاً تثبیت شده و پیشرفته را روی مجموعه داده خود اعمال می‌کنیم و مجموعه‌ای از نتایج پایه را بر اساس پژوهش سایر محققان ایجاد می‌کنیم تا با آن‌ها محک زده شود. در مجموعه داده‌ها با توجه به دانش ما، دو مجموعه داده موجود برای پاسخگویی به سوالات دیداری وجود دارد. مجموعه داده VQA روی حدود ۵۰۰۰ عکس ایجاد شده و دارای حدود ۳۰۰۰۰ جفت سوال-پاسخ است. سوالات واقعاً چالش برانگیز هستند. بااین حال، فقط تعداد معدودی سوال از این دست وجود دارد. مجموعه داده VQA بر روی تصاویر واقعی در MS COCO و تصاویر انتزاعی ساخته شده‌است. قراراست در ادامه پروژه جفت‌های پرسش و پاسخ منطبق با اقدامات انسانی ایجاد می‌شود که می‌تواند به پرسش‌های "جالب" و "متنوع" تقسیم شوند. روند پردازش بر مجموعه داده VQA گسترش می‌یابد تا با دستیابی به تعادل بیشتر بین اطلاعات دیداری و متنی، با جمع آوری تصاویر مکمل، به روشی که هر سوال با یک جفت تصویر مشابه با پاسخ‌های مختلف همراه باشد، حاصل گردد. در مجموعه داده‌های COCO، جفت سوال-پاسخ به طور خودکار از زیرنویس‌های تصویر براساس تجزیه نحوی و قوانین زبانی تولید می‌شود. دراینجا تولید خودکار جفت‌های پرسش و پاسخ مجموعه داده‌های موجود، از روش‌های خودکار برای ساخت جفت سوال-پاسخ استفاده کرده‌اند. همچنین در COCO، الگوریتم پایه تولید کننده سوال بر اساس تجزیه نحوی است و می‌تواند چهار نوع سوال را تشکیل دهند، که شامل "شی"، "تعداد"، "رنگ" و "مکان" است و با استفاده از قوانین ساده ساخته می‌شوند. برای آموزش چنین مدلی، به جفت تصویر-سوال نیاز است و با مشکل مرغ و تخم مرغ روبرو می‌شود. دراینجا هدف تولید سوال است، اما تحقق این هدف به سوالات ایجاد شده نیاز دارد. در دیتاست VQA، تصاویر را همراه با اطلاعات جانبی مرتبط (به عنوان مثال، زیرنویس‌ها، روش‌ها، صفحات) از پایگاه داده انتخاب کرده و بر اساس الگوهای تعریف شده در جفت سوال و پاسخ ایجاد می‌کنیم. تولید سوال دراین بخش، در مورد چگونگی تولید خودکار سوالات از زیرنویس‌ها حاصل می‌شود. شکل۱ چارچوب کلی الگوریتم را نشان می‌دهد. ما پردازش را بر اساس زبان برنامه نویسی پایتون انجام می‌دهیم. که در آن مواردی مانند تقسیم جمله، رمزگذاری، برچسب گذاری بخشی از گفتار (POS)، شناسایی موجودیت (NER) و تجزیه تشکیل دهنده، تشکیل می‌شود.



شکل (۱) روند شماتیک الگوریتم

سیستم‌های VQA تعاملات لازم با سیستم‌های مدیریت اطلاعات را باز تعریف می‌کنند. در بازیابی اطلاعات به شکل سنتی آن، کاربران سیستم‌های اطلاعاتی را با کلمات کلیدی جستجو می‌کنند تا فهرستی از اسناد سازگار را به دست آورند. با این وجود مرحله دوم ضروری به نظر می‌رسد که در آن کاربر باید پاسخ را از یک سند/تصویر خاص استخراج کند. برعکس، سیستم‌های پرسش و پاسخ این امکان را برای کاربران فراهم می‌کنند که سوال خود را مستقیماً به زبان طبیعی بیان نموده و همچنین پاسخ را با زبان طبیعی بازیابی کنند. زیربنای این رویکرد غالباً یک فرایند دو مرحله‌ای است که در آن سیستم پرسش و پاسخ ابتدا سند مورد نظر را از صف مربوطه شناسایی می‌کند و متعاقباً جواب صحیح را از یک سند استنباط می‌کند. در نتیجه، پاسخ به سوالات مسیری را برای سیستم‌های اطلاعاتی ارائه می‌دهد که می‌تواند به سهولت استفاده کمک شایانی کند. دوم اینکه، سیستم‌های پرسش و پاسخ قول تسریع در روند جستجو را می‌دهند، زیرا کاربران مستقیماً پاسخ صحیح به سوالات خود می‌گیرند. در عمل، این فرآیند میزان زیادی از عملیات و زمان خواندن غیر خودکار را که برای شناسایی سند مربوطه و قرار دادن اطلاعات مناسب در داخل یک مورد لازم است، کاهش می‌دهد.

یکی از برجسته‌ترین سیستم‌های فعلی پرسش و پاسخ IBM Watson است. تلاش‌های تحقیقاتی بیشتر در زمینه پاسخگویی به سوالات، منجر به‌استفاده از سیستم‌هایی برای برنامه‌های کاربردی و عملکرد سیستم‌های پرسش و پاسخ فعلی در تنظیمات دنیای واقعی می‌شود که‌اغلب محدود است و در نتیجه رضایت کاربر را کاهش می‌دهد. دراینجا ایده انتقال دانش از یک برنامه جنرال و حوزه باز به مورد استفاده از دامنه خاصی است. این روش بسیار مقرون به صرفه است زیرا صرفاً به مجموعه کوچکی از چند صد جفت سوال پاسخ دارای برچسب برای تنظیم دقیق طبقه بندی‌های یادگیری ماشین و با برنامه‌های دامنه خاص نیاز دارد. علیرغم این واقعیت که فرا داده‌ها به طور معمول در پایگاه‌های دانش استفاده می‌شود، همچنین ما از

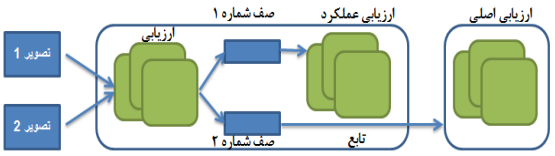
موارد استفاده شده قبلی در سیستم‌های مبتنی بر محتوا برای پاسخ به سوالات سیستم‌های عصبی آگاهی نداریم.

بر اساس تجربه عملی، ما دو اهرم مهم را که در متناسب کردن سیستم‌های پرسش و پاسخ به برنامه‌های خاص کمک می‌کند، شناسایی می‌کنیم. یعنی یک فیلتر فرا داده[1] ثانویه که به عنوان ابزاری برای استفاده مجدد از دانش محدود شده‌است، استفاده می‌شود. به دیگر سخن، دامنه اسناد انتخاب شده برای استخراج پاسخ و یادگیری انتقال است. توابع پاسخ به سوالات در دامنه باز این مکانیسم‌ها، اجزای مختلف یک سیستم پرسش و پاسخ را هدف قرار می‌دهند. فیلتر Meta Data بر رفتار ماژول‌ها تأثیر به سزایی می‌گذارد، در حالی که سیستم یادگیری به‌استخراج پاسخ می‌پردازد. دانش انباشته در سیستم‌های پرسش و پاسخ مبتنی بر محتوا توسط مجموعه اصلی اسناد آن تشکیل شده‌است. در حالی که در اکثر برنامه‌ها، سیستم خود محدوده وسیعی از دانش را پوشش می‌دهد، اما اسناد جداگانه فقط جنبه خاصی ازاین دانش را برطرف می‌کنند.

تحلیل نتایج

بر خلاف تنظیمات رایج یادگیری تحت نظارت که در آن عملکرد در یک مجموعه آزمون استاتیک سنجیده می‌شود، ارزیابی عملکرد گسترش یافته این عوامل که توالی اجرا را در محیط آزمایش تغییر می‌دهند کار چندان ساده‌ای نیست. ارزیابی این عوامل شامل اجرای کد کاربران بر روی مجموعه‌ای از محیط‌های دیده نشده‌است که یک مجموعه آزمایش‌های مخفی را تشکیل می‌دهد. به طوری که می‌توان الگوریتم‌های مختلفی را بیش از حد در محیط‌های آموزشی مورد تجزیه و تحلیل قرار داد. این مکانیزم یک سیستم عامل منبع باز بسیار قابل گسترش است که نیازهای اساسی را برآورده می‌کند. جامعه هوش مصنوعی برای ارزیابی انسان، مبتنی بر چارچوب مدل‌های یادگیری ماشین است.

در یک محیط پویا به جای ارسال ساده پیش بینی‌ها در مجموعه داده‌های استاتیک، امکان ارزیابی عوامل به صورت تعاملی به میزان مناسبی وجود دارد. با اجرای آزمایش‌های مربوط به‌این پیش بینی‌ها، نتایج شگفت انگیزی یافت می‌شود، که به نظر نمی‌رسد عملکردی که از طریق روال‌های عادی بدست آمده‌است در این مجموعه تعمیم یابد.

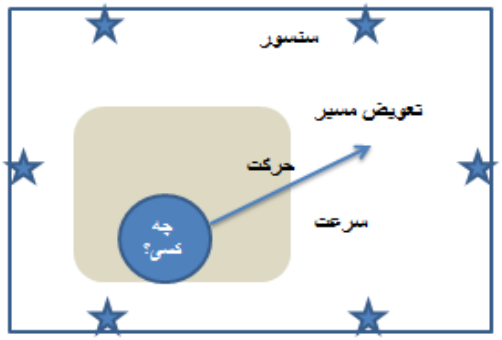


شکل (۲) چرخه‌ارزیابی عملکرد

برای تشخیص حرکت انسان با خروجی‌های دیجیتال این حسگرها به طور خاص با استفاده از رویکردهایی جدید در حال کار بر روی شناسایی جهت حرکت و شمارش افراد ورودی یا خروجی از ورودی اتاق یا ساختمان با استفاده از سیگنال خروجی خاموش سنسورها هستند.

در این مکانیزم‌ها تا ۹۹٪ صحت تشخیص جهت حرکت به عقب و جلو و ۹۵٪ دقت تشخیص تعداد عابران را نشان می‌دهند. در عین حال به سادگی با مشاهده اختلاف زمان بین سنسورهای رو به داخل / خارج، جهت حرکت را در یک گیت متمایز می‌کنند. همچنین رویکردی برای تخمین تعداد افراد در هر فضای اداری با استفاده از جفت‌های توزیع شده سنسورهای کنترل و الگوریتمی برای پردازش اطلاعات حسگر توزیع شده پیشنهاد شده‌است.

تشخیص حرکت انسانی با خروجی‌های آنالوگ نیز جز روش‌های جدیدی برای تشخیص جهت حرکت برای یک جسم در حال حرکت در میدان دید یک سنسور منفرد است، که عناصر دوگانه حسگر آن بطور معکوس به صورت قطبی طراحی شده و در صفحه حرکت سنسور قرار می‌گیرند.



شکل (۳) روند تشخیص عملکرد

با تشریح نیازها در یک بستر ارزیابی که بتواند وظایف پیچیده و فزاینده‌ای را که توسط جامعه آماری برطرف می‌شود به خوبی ارزیابی کند، به صراحت موارد قابل آزمایش را مشخص می‌کنیم که یک ابزار ارزیابی مدرن باید آن‌ها را برآورده کند. با تمرکز بیشتر بر جامعه آماری در مسائلی که سعی در حل آن‌ها داریم، متوجه تغییر از مجموعه داده‌های ایستا به محیط‌های پویا شده‌ایم. به‌این ترتیب، تمامی‌سیستم عامل‌های ارزیابی مدرن باید بتوانند عوامل ارسالی را در این محیط‌ها اجرا کنند. یک پلتفرم ارزیابی نیاز به پشتیبانی از تعداد دلخواه فازها و انشعابات داده برای تأمین کردن جامعه آماری دارد. این مهم اغلب از چند تقسیم مجموعه داده استفاده می‌کنند و هرکدام دارای اهداف متفاوتی هستند. با جدا کردن صفحه‌اصلی از گره‌های آماری که به طور واقعی ارزیابی را انجام می‌دهند، سیستم عامل باید امکان میزبانی فرایندهای ارزیابی را به طور مستقل در خارج از سیستم فراهم کند. با اجازه دادن به عوامل خارجی، برگزارکنندگان چالش این امکان را دارند که محاسبه‌اضافی برای پاسخگویی به نیازهای مقیاس پذیری یک چالش جدید ارائه دهند.

در این سلسله تحقیق، به طور مفصل ویژگی‌های اصلی ارائه شده توسط الگوریتم ارزیابی مبتنی بر هوش مصنوعی مانند پروتکل‌ها و مراحل ارزیابی مشتری، ارزیابی از راه دور و رابط خط فرمان توصیف می‌شود. همچنین چرخه حیات یک چالش به عنوان مبنا توضیح داده خواهد شد. سپس، در مورد چگونگی ارزیابی غیر همگام صحبت خواهد شد. یکی از اهداف ما گسترش مطالعات در مورد دو چالش مختلف میزبانی شده در این الگوریتم است که در این زمینه بحث می‌کنیم. یکی از مهمترین چالش‌ها پاسخ به سوالات به صورت دیداری است، جایی که عملکرد الگوریتم چالش دیگری است که در چالش پاسخ سوالات دیداری مقایسه می‌کنیم. در اینجا کار خود را در مورد ارزیابی عوامل مکالمه بصری از طریق نمایش عدم تطابق بین محک زدن AI به طور جداگانه ادامه می‌دهیم.

در این مکانیزم از روی اجبار چندین چالش مرتبط با ایجاد چارچوبی برای ارتباط فرآیندها با عوامل اجرایی را کنار می‌گذاریم. اما برای اطمینان از

[1] Meta Data

اینکه الگوریتم به صورت ناقص ارزیابی مجدد نشده و فواصل می‌توانند مجدداً با همان عامل متصل شوند (اگر اتصال فقط به طور موقت قطع شود) حسابداری زیادی انجام می‌دهیم.

برای حل این مشکل، ما به سازمان دهندگان چالش اجازه می‌دهیم خوشه گره‌های فرزند را برای ارزیابی بازفراخوانی کنند، در حالی که سیستم از چالش‌های میزبانی، مدیریت ارسال‌های کاربر و حفظ تابلوی امتیازات مراقبت می‌کند. سیستم پیشنهادی ما با استفاده از صف‌های پیام، گره‌ها را از وب سرورها جدا می‌کند. محموله پیام شامل تمام اطلاعات لازم برای اجرای ارزیابی در مورد ارسال آن به صفحه اصلی است.

چهار بخش اصلی از زیرساخت‌ها وجود دارد: چارچوب شناسایی حلقه‌ها بر اساس سرویس Tocker و Elastic Container Service وظیفه مقیاس گذاری، استقرار و نظارت بر زیرساخت‌ها را برآورده می‌کند تا نیازهای همه چالش‌های موجود در سیستم را برآورده کند. وب سرورها و پایگاه داده‌های مبتنی بر VQA وCOCO، APIهای کاملی را برای بررسی چالش و شرکت کنندگان در تعامل با سرویس ارائه می‌دهند. برای مولفه‌های خاصی که از خدمات اختصاصی ارائه دهنده ابر خاص استفاده می‌کنند، سعی شده‌است تا پروتکل با گزینه‌های منبع باز سازگار باشد.

جدول(۱) اطلاعات دیتاست‌ها

نوع پاسخ	جفت‌ها QA	تعداد تصویر	دامنه	دیتاست ها
باز	۱۲۵۰۰	۱۴۴۹	عمومی	DAQUAR
باز/محدوده	۶۰۰۰۰۰	۲۰۰۰۰۰	عمومی	VQA
باز/محدوده	۱۲۰۰۰۰۰	۲۰۰۰۰۰	عمومی	VQA v2
باز/محدوده	۱۱۸۰۰۰	۱۲۰۰۰۰	عمومی	COCO-QA
باز/محدوده	۱۵۰۰۰	۴۰۰۰	پزشکی	VQA-Med
باز/محدوده	۳۵۰۰	۳۰۰	پزشکی	VQA-RAD

در یک مجموعه داده پاسخ سوال بصری (VQA) شامل ۵۰۰۰ تصویر و ۳۰۰۰۰ جفت سوال-پاسخ ایجاد می‌کنیم تا تحقیقات VQA پزشکی را تقویت کنیم. از نظر دانش ما، این اولین مجموعه داده برای آسیب شناسی VQA است. ما یک خط لوله نیمه اتوماتیک ایجاد می‌کنیم تا بتواند به طور موثر مجموعه داده‌های پزشکی VQA را از کتاب‌های درسی پزشکی و کتابخانه‌های دیجیتال آنلاین ایجاد کند. خط لوله ما میتواند به طور گسترده‌ای در سایر حوزه‌های تصویربرداری پزشکی فراتر از آسیب‌شناسی اعمال شود، مانند رادیولوژی، سونوگرافی و غیره.

پیکربندی مجموعه‌ای از پارامترهای مفید از جمله عنوان چالش، تاریخ شروع، تاریخ پایان، تعداد مراحل چالش، تعداد ارسال‌های مجاز در روز برای هر یک از این مراحل همراه با چندین مرحله دیگر است. علاوه بر این، ما به کاربران اجازه می‌دهیم جزئیات چالش را به شکل الگوهای HTML اضافه کنند تا در صورت اجرای مسابقه در صفحه وب نمایش داده می‌شوند. از این موارد می‌توان برای معرفی کار، توضیح پروتکل ارزیابی و توصیف مجموعه داده برای شرکت کنندگان استفاده کرد. در مجموعه برنامه‌نویسی تهیه شده یک نمونه راهنما آورده شده‌است تا برخی از زمینه‌های مهم را فهرست کند از جمله اسکریپت‌های ارزیابی مرتبط با چالش، قسمت مهم بعدی بسته را تشکیل می‌دهد.

در این چالش می‌توان تصویر را در بالای یک تصویر پایه که ارائه داده می‌شود و شامل تمام قطعات لازم برای اتصال اسکریپت ارزیابی با ارسال‌های ورودی در تولید است، ساخت. بسته به میزان چالش، موارد آزمایش می‌توانند دو نوع استعلام ارسال کنند که شامل تمام پیش بینی‌ها در یک مجموعه آزمون استاتیک ارائه شده توسط سازمان دهندگان چالش یا یک تصویر حاوی مدل پایه است که بعداً برای ارزیابی سناریوهای آزمون جدیدتر مورد استفاده قرار می‌گیرد. ارسال یک بسته شامل بارگذاری یک پیام در قالب تعیین شده در سرور ارزیابی است. هنگامی‌که یک شرکت کننده یک درخواست را برای یک چالش ارسال می‌کند، تابع فراخوانی می‌شود که ابتدا ارسال را در یک پایگاه‌داده ذخیره می‌کند و پیامی را به همراه متادیتای ارسال به صف منتشر می‌کند. پیام با استفاده از کلید مسیریابی مربوط به چالش به یک صف خاص منتقل می‌شود. کاربران همچنین می‌توانند عوامل خود را برای ارزیابی پویا یا در محیط‌های شبیه سازی از طریق AMT یک شرکت کننده که آزاد است با نصب وابستگی‌های اضافی، تصویر Tocker را گسترش دهند. با جدا کردن عکس فوری مدل و تصویر حاوی کد برای اجرای عامل، ارسال‌های بعدی سریع تر خواهد شد زیرا شرکت کنندگان مجبور نیستند بارها و بارها تصاویر Tocker را در رجیستری بارگذاری کنند. هنگام ارسال، درست زمانی که گره عامل آماده‌ارزیابی شرکت کننده باشد، یک بسته عامل را بر اساس تصویر ارسالی راه‌اندازی می‌کند و مدل و محیط آزمایش را به عنوان حجم به بسته پیوست می‌کند. عامل برای شروع ارزیابی کار، متن "ارزیابی شده" را درون بسته اجرا می‌کند. امتیازها و معیارهای اصلی توسط عامل به تابلوی امتیازات مرکزی ابلاغ می‌شود.

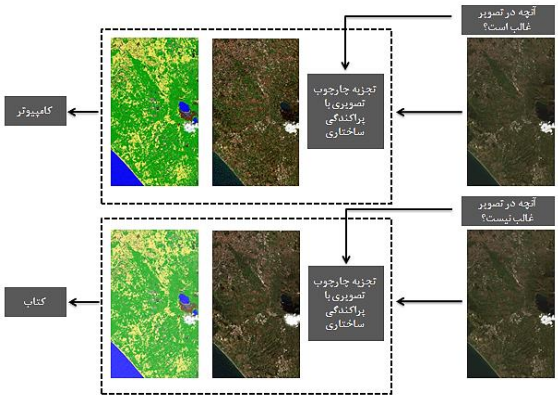
مجموعه کد نگارش شده در قالب شبیه سازی به نام Human VQA توسعه داده شده‌است که (تا آنجا که می‌دانیم) بر اساس آخرین وضعیت موجود در مجموعه داده VQA1 و COCO استوار است.

تابع اصلی در مورد مدلهای آموزشی بر پایه End to End در یک مجموعه داده چند حالته استوار است که از سه گانه زیر ساخته شده‌است:

- تصویری که هیچ اطلاعات دیگری به جز پیکسل‌های اولیه ندارد
- سوالی درباره محتوای بصری در تصویر مرتبط اجرا گردد
- پاسخ کوتاه به سوال (در یک یا چند کلمه)

همانطور که در تصویر ۴-۴ مشاهده می‌کنید، دو سه گانه مختلف (بر روی همان) از مجموعه دادهVQA نشان داده شده‌است.

این مدل‌ها نیاز به یادگیری نمایش‌های چند حالته بسیار قوی دارند تا بتوانند پاسخ‌های صحیح ارائه بدهند.



شکل ۴) نمایش پردازش تصویری در پیاده سازی

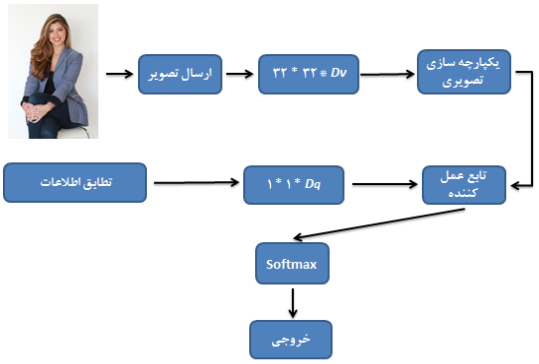
تابع VQA هنوز روی تحقیقات فعال است. با این حال، وقتی این مشکل حل شد، بهبود رابط‌های انسان به ماشین بسیار مفید است. بینش سریع در مورد روش پیاده سازی شده جامعه آماری VQA و COCO روشی مبتنی بر چند مولفه قابل یادگیری تدوین شده‌است: بینش صریح در مورد روش ما جامعه VQA رویکردی بر مبنای چهار مولفه قابل یادگیری دارد:

- یک مدل سوال که می‌تواند یک مدل LSTM، GRU، یا منطق پیش یادگیری باشد.
- یک مدل تصویر که می‌تواند یک مدل از Res-Net152 باشد.
- یک طرح ترکیبی که می‌تواند شامل مدل‌هایMCB، MLB، یا Human باشد.
- در صورت لزوم یک طرح آگاه کننده که ممکن است چندین دیدگاه را به همراه داشته باشد.

مهم‌ترین دلیل این است که ادغام چند حالته بین تصویر و بازنمایی سوال یک مولفه مهم است. بنابراین، مدل پیشنهادی ما با استفاده از تجزیه و تحلیل تاکر از حسگرهای رایج همبستگی برای مدل سازی تعاملات چند حالته دقیق‌تر، به منظور ارائه پاسخ مناسب بهره می‌برد.

بهترین مدل از نگاه ما بر اساس:

۱) تفکرات قبلی آموزشی برای مدل سوال
۲) ویژگی‌های یک مدل آموزشی با تصاویر استاندارد برای مدل تصویر
۳) مکانیزم Human پیشنهادی ما (بر اساس تجزیه تاکر) برای طرح همجوشی
۴) یک طرح اعلان کننده با دو View کلی



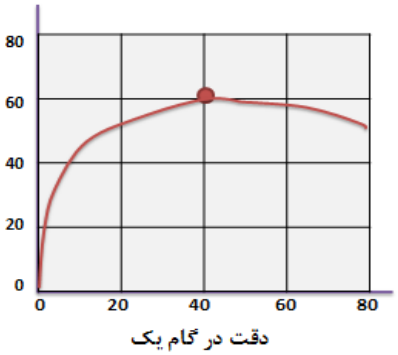
شکل (۲) چرخه پردازش تصویری در پیاده سازی

در حال حاضر ۲ دیتاست ارائه شده‌است:

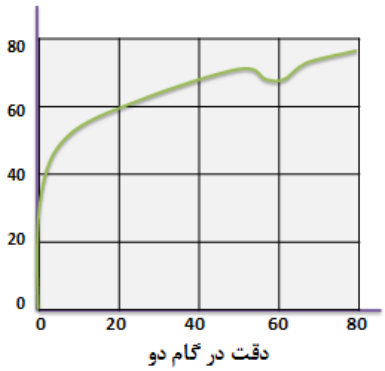
- COCO
- VQA

که درحال‌حاضر برای استخراج ویژگی‌ها استفاده می‌شوند. تصاویر مورد نیاز به صورت خودکار در data-dir بارگیری می‌شوند و ویژگی‌ها به طور پیش فرض با استاندارد استخراج می‌شوند. حداقل فضای تصویری برای بخش نمایش می‌تواند ۲۰۴۸*۱۴*۱۴ باشد. همچنین ما برای ارزیابی دیتا از یک محیط اجرایی با پردازنده Corei۷ و فضای حجیمی در SSD استفاده کردیم، از جمله فضای ۳۲ گیگابایتی برای تصاویر، ۱۲۵ گیگابایت برای ویژگی‌های جریان پردازش و ۱۲۳ گیگابایت برای ویژگی‌های آزمون و ۶۱ گیگابایت برای ویژگی‌های اجرایی.

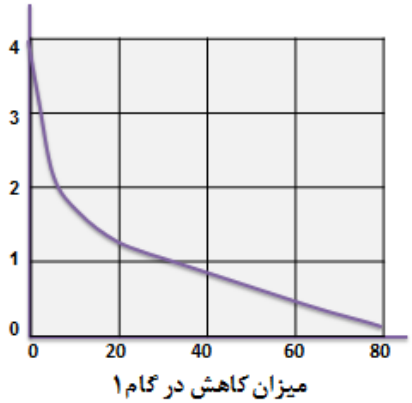
با استفاده از برنامه ریزی برای نظارت بر آموزش، یک تجسم بسیار عالی از آزمایش ایجاد می‌شود. توجه داشته باشید که باید مشاهده کنید تا اولین بار عامل اجرایی پردازش شود تا مکانیزم به پایان برسد و سپس فایل html ایجاد می‌شود و در مرورگر پیش فرض ظاهر می‌شود. مکانیزمhtml هر ۶۰ ثانیه بروز رسانی می‌شود. با این حال، در حال حاضر باید با F۵ در مرورگر برای رفرش استفاده نمایید.



شکل (۳) نمودار دقت (Accuracy) در گام یک



شکل (۴) نمودار دقت در گام ۲

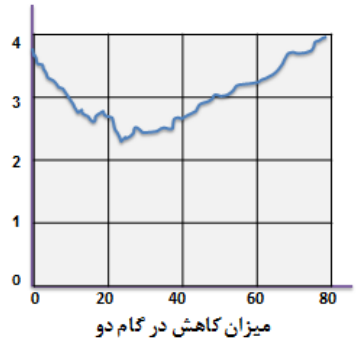


شکل (۵) نمودار کاهش در گام ۱

وقتی روند اجرایی زنده شد، سیستم دایرکتوری حاوی مجموعه داده را در محفظه ارزیابی بار می‌کند و عاملی را که به صف برای گرفتن مطالب ارسالی ورودی گوش می‌دهد، اداره می‌کند. محموله دریافت شده از صف پیام حاوی کلیه فرا داده‌های مفید مانند مسیر برای ارزیابی انسان در چرخه حیات است. کد ارزیابی ابتدا عامل را بارگیری نموده و یک HIT[1] جدید در سیستم راه‌اندازی می‌کند. هنگامی‌که عامل HIT پذیرفته می‌شود، عامل با رابطی که

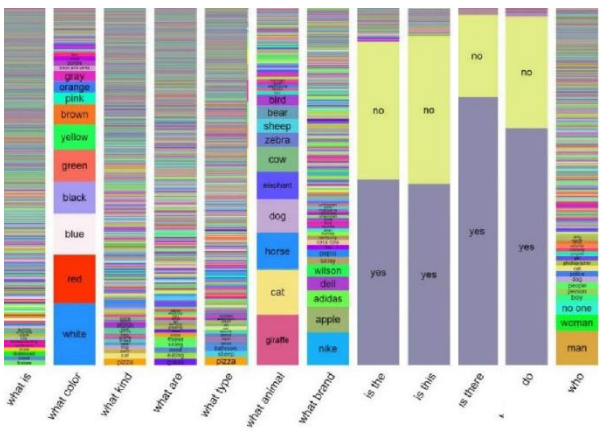
[1] Human Intelligence Task

در داخل یک تصویر Tocker در حال اجرا است هماهنگ می‌شود. بر اساس دستورالعمل ارائه شده، عامل با رابط تعامل خواهد کرد و آن را طبق معیارهای خاص ارزیابی می‌کند. داده‌های واکنش متقابل و رتبه بندی داخلی داده شده توسط عامل ذخیره می‌شود که در نهایت بر روی تابلوی امتیازات تخلیه می‌شود. برنامه نگارش شده مسئولیت مدیریت اتصال مداوم بین نماینده و عامل، مدیریت خطا، تلاش مجدد، ذخیره داده‌های تعامل متناظر با HIT و تأیید یا رد خودکار را بر عهده دارد. با آماده نگه داشتن لیست عامل‌ها برای ارزیابی در راستای ارسال جدید، از تأخیر ناشی از بارگیری مجدد کد ارزیابی بارها و بارها جلوگیری می‌شود.



شکل (۹) نمودار میزان کاهش در گام ۲

با تمرکز بر گفتگوی و پرسش و پاسخ تصویری، جایی که به یک تصویر، یک سابقه گفتگوی مرتبط و یک سوال پیگیری در مورد تصویر داده می‌شود، یک عامل لازم است که به سوال پاسخ داده شود. در حالی که سوال در تاریخ خاصی ایجاد می‌شود. از آنجایی که ارزیابی بیشتر با مجموعه عظیمی از احتمالات پیچیده‌تر توام است، پاسخ‌های صحیح و نمونه برداری نسبتاً پراکنده از این فضا، حتی در مجموعه داده‌های مقیاس بزرگ نیز امری ضروری به نظر می‌رسد. با توجه به‌این سختی‌ها و ماهیت تعاملی کار، روشن است که مناسب‌ترین راه برای ارزیابی عوامل گفتگو، استفاده از یک عامل در حلقه‌است، یعنی چیزی به مانند یک آزمون تورینگ بصری. پس از ۱۰ دور تعامل انسان و عامل، رتبه بندی عامل از انسان به صورت یک امتیاز در جدول واقعی روش ما در زمان واقعی کاهش می‌یابد. سرانجام، هنگامی‌که ما به سمت توسعه عوامل هوشمند برای کارهایی که در محیط‌های فعال به جای مجموعه‌های داده ساکن قرار دارند، حرکت می‌کنیم، جایی که عوامل برای تغییر وضعیت کلی پیرامون خود اقدام می‌کنند، ضروری است که‌ابزارهای جدیدی برای ایجاد دقیق معیارها در محیط‌ها بسازیم.



شکل (۱۰) پاسخ‌های نگاشت شده بر تصاویر

شکل پاسخ‌های نگاشت شده بر تصاویر: بیشتر پاسخ‌ها از یک کلمه تشکیل شده‌است، توزیع پاسخ‌ها شامل یک، دو یا سه کلمه است که به ترتیب ٪۳۲.۸۹، ٪۹۱.۶ و ٪۷۴.۲ برای تصاویر واقعی و ٪۵۱.۹، ٪۸۹.۵ و ٪۴۹.۲ برای صحنه‌های رفتار انسانی و اختصار پاسخ‌ها تعجب آور نیست، زیرا سوالات معمولاً اطلاعات خاصی را از تصاویر استخراج میکنند.

شناسایی فعالیت انسانی برای دهه‌ها به طور کامل مورد تحقیق قرار گرفته و هیچ کمبودی نداشته‌است. درواقع مستند سازی رویکردهای مختلف به‌این کار از اهداف شبیه‌سازی است، یعنی:

- چرخش بیشتر در اطراف سیگنال پردازش
- ویژگی‌های دستی که برای مدل‌های یادگیری ماشین مرسوم مورد استفاده قرار می‌گیرند.

اخیرا، با پیشرفت شبکه‌های عصبی مصنوعی که پیش از این برای عملکرد مناسب خود به رسمیت شناخته شده‌بودند، در پردازش تصاویر کامپیوتری، کاربردهایی در مدل‌سازی دنباله نیز پیدا شده‌است.

شبکه‌های عصبی مصنوعی(RNN)[1] و به طور خاص شبکه‌های حافظه بلند مدت (LSTM)[2]، دقت زیادی به کارهای مدل‌سازی که در آن ورودی‌ها از یک دنباله هستند، به دست آورده‌اند که با گذر زمان از یک شبکه LSTM تغذیه می‌شوند. توانایی شبکه‌های عصبی برای مدل سازی مجدد و ویژگی‌های موقتی محاسبه شده از طریق زمان به آن‌ها مزیت خاصی در مدل‌سازی می‌دهد. با توجه به اینکه فعالیت‌های انسانی به شدت به فعالیت‌های قبلی مربوط است، در اجرای ساده، از یک دیتاست CNN به عنوان مدل پایه، به صورت یک نقطه شروع استفاده می‌کنیم.

فرض کنید اگر دیتاست پرسش‌ها مبتنی بر بخش تصویر-پردازش پزشکی بود، می‌توانیم از جدول نمونه زیر استفاده کنیم:

جدول (۲) پرسش‌ها مبتنی بر بخش تصویر-پردازش پزشکی

نوع	سوال	جمله‌اصلی
چیست	چه چیزی در منطقه اپی فیز پخش شده‌است؟	انتهای استخوان بلند در ناحیه‌اپی فیز پخش شده‌است.
کجا	در این نمای چهار کانال قلب بطن چپ کجاست؟	بطن چپ در پایین راست قرار دارد.
چه وقت	در طی چه مدت زمانی اکثر زخم‌ها بهبود میافته‌اند؟	بعد از ۱ سال مراقبت. بیشتر جای زخم‌ها از بین رفته‌است.
چه تعداد	چند سنگ صفراوی چند وجهی در بخش لومن وجود دارد؟	دوسنگ صفرا چند وجهی در لومن وجود دارد.
چه کسی	سلول‌های تومور و هسته‌های آن‌ها از جنبه ظاهری کاملا یکنواخت است؟	سلول‌های تومور و هسته‌های آن‌ها از لحاظ ظاهری کاملا یکنواخت است.
چگونه	استخوان ترایکولار چگونه عمل کرده و آیا ترایکول‌ها فضای مغز رانشان میدهد؟	استخوان ترایکولار که فضای مغز را تشکیل میدهد و ترایکولار فعالیت پوکی استخوان در حاشیه را نشان میدهد.

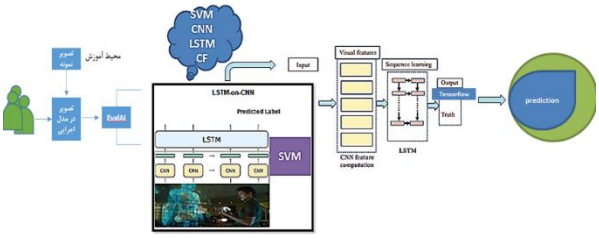
[1] Recurrent neural network
[2] Long short-term memory

همچنین با توجه به وضعیت دیتاستها و روش کاربردی، آمار زیر را می‌توان ارائه داد، که این اعداد براساس عملیات اجرایی حداقل و حداکثر تعداد سوالات که میتواند بر هر تصویر نگاشت شود ارائه می‌شود.

جدول (۳) آمار نسبی دیتاستها

*	حداکثر	میانگین	حداقل
سوالات بر تصویر	۱۵	۶.۲	۱
کلمه بر سوال	۲۹	۹.۲	۴
کلمه بر پاسخ	۱۱	۲.۳	۱

با توجه به یک تصویر و یک سوال زبان طبیعی در مورد تصویر، وظیفه اصلی روش ما ارائه پاسخی دقیق با زبان طبیعی است. سوالات بصری بخش‌های مختلف یک تصویر را انتخاب می‌کنند، از جمله جزئیات پس زمینه و پیش زمینه. با توجه به یک تصویر و یک سوال زبان طبیعی در مورد تصویر، وظیفه روش ما ارائه پاسخ دقیق زبان طبیعی است. در روش‌های قبلی، از طبقه‌بندی فعالیت‌های انسانی از طریق اندازه‌گیری‌های سنسور با داده‌های کمی و تا جای ممکن مهندسی شده استفاده می‌شد. در برخی از پردازش‌ها هر واحد اندازه‌گیری نرمالیزه می‌شد. خواندن را در یک دسته از مجموعه داده‌ها انجام داده و این خواندن به طور مستمر و نه به صورت دسته‌ای ادامه پیدا می‌کرد و یک زمان اجرای هر مرحله به شکل واحد در یک بردار ۴۰ x ۱ به عنوان ورودی استفاده می‌شد. این بردار ویژگی با استفاده از انواع روش‌های ماشین لرنینگ به ویژه از طریق شبکه‌های عصبی با چندین لایه کانولوشن و ۲ لایه به طور کامل متصل می‌شدند که هر یک موجب ایجاد مشکلاتی در افزایش دقت سیستم می‌شدند که در روش پیشنهادی ما به میزان قابل توجهی رفع شده‌است.



شکل (۱۱) چرخه فعالیت سیستم مورد تحقیق

در کل ۱.۱ میلیون سوال و ۲۵۰۰۰ تصویر هر سوال با متن یک تصویر خاص ساخته شده‌است. اما جا دارد به طور خلاصه درباره VQA به نکته‌ای اشاره کنیم که این سیستم به پاسخ سوالات بصری مبتنی بر دانش اولیه یادگیری ماشین نیاز دارد که در اینجا مدل پیشنهادی از دو معماری وابسته تشکیل شده‌است:

- مدل CNN برای پردازش تصویر
- مدل LSTM برای پردازش سوال

لازم به ذکر است ما خروجی نهایی آموزش و تست برنامه پیشنهادی خود را که در سیستم با پردازنده مناسب راه‌اندازی گردیده‌بود، در سیستمی با پردازنده i5 و روی بستر پایتون با مدت زمان اجرایی نسبتا طولانی اجرا کرده و خروجی نهایی دقت روش پیشنهادی خود را به دست آوردیم.

مدل LSTM

مبنای خروجی از طریق لایه‌های شبکه عصبی کاملاً متصل (MLP)[1] ارسال می‌شود که بر وضوح ۱۰۰۰ * ۱۰۰۰ و ۱۰۰۰ * ۱۰۲۴ است. در مدل LSTM دو لایه استفاده شده و با ابعاد پنهان ۵۱۲ بعد شامل کلمات ورودی تعبیه شده می‌شود، که با استفاده از دیتا تعبیه شده GLOVE خروجی از ابعاد ۲۰۴۸ * ۵۱۲ (۲ حالت پنهان) ۲ * ۵۱۲ (۲ حالت سلول) عبور از لایه کاملاً متصل (۱۰۲۴ * ۲۰۴۸) را شامل می‌شود[19].

مدل CNN

خروجی نهایی حاصل از لایه خروجی پس از بررسی توسط soft max، مدل خاص CNN VGG و خصوصا لایه‌های کانولوشن خروجی تولید شده از ابعاد ۴۰۹۶ و عبور از لایه کاملاً متصل است[19].

نتایج قادر به پیش بینی پاسخ صحیح با دقت ~ ۴۰٪ در پیش بینی پاسخ‌های ساده تر مانند:

- بله، خیر
- ۳، ۴ و ۵
- آبی، قرمز و سبز

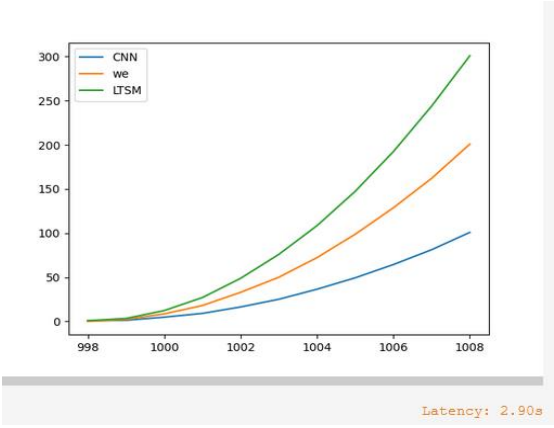
در روش پیشنهادی، ما برای مقایسه با دو الگوریتم منتخب شبکه‌های عصبی (ماشین لرنینگ) یعنی CNN و LSTM در ابتدا محدوده‌های رصد خود را بر دسته‌های ۵۰ تایی خوشه‌ها (محور عمودی) و تعداد نگاشت‌ها بر تصاویر (محور افقی) منطبق می‌نماییم که نمودار ۱ حاصل می‌شود. که در آن روش پیشنهادی ما مقیاس بالانس شده را نمایش می‌دهد (در عین حال به صعود نمودارها به صورت تقریبا یکسان توجه میکنیم. در مرحله بعدی تعداد خوشه‌ها را به ۱۰۰۰ تایی و تعداد نگاشت‌ها را نیز افزایش میدهیم که رشد نمودارها در کل یکسان است. (نمودار شماره ۲)

در مرحله پیاده‌سازی با استفاده از خوشه‌های ۱۰۰ تایی از مجموعه داده‌ها به ترتیب مراحل زیر اجرا می‌شود:

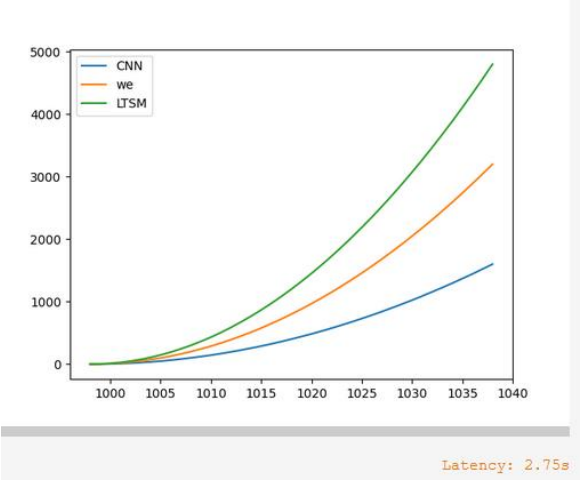
۱. میانگین متوسط اجرای هر سه روش CNN، روش ما و محاسبه LTSM و در سه تایی‌های مرتب نمایش داده می‌شود.
۲. میانگین پاسخگویی هر سه روش با ۱۷ رقم اعشار ظاهر می‌شود.
۳. میانگین متوسط به‌ازا الگوریتم کمکی CNN مبتنی بر دیتاست VQA1 نمایش داده می‌شود.
۴. میانگین متوسط به‌ازا الگوریتم کمکی LTSM مبتنی بر دیتاست VQA1 نمایش داده می‌شود.
۵. سپس یک بار دیگر میانگین پاسخگویی هر سه روش با ۱۷ رقم اعشار ظاهر می‌شود.
۶. دقت روش ما محاسبه می‌گردد.
۷. میزان سنجش نرمالیزه شده سیستم محاسبه می‌شود.
۸. میزان مسئولیت پذیری سیستم نمایش داده می‌شود.

سپس با استناد به محاسبات سیستمی۸ گانه فوق دقت ماشین را در محدوده ۰.۰۵ تا ۰.۸ به شکل یک آرایه یک بعدی نمایش می‌دهد و ۰.۰۵ یعنی خروجی برنامه کاملا نامعتبر است و ۰.۸ به معنی درستی حد اکثری برنامه با ۸۰ درصد دقت است و نیز بازه بیشتر از ۸۰ درصد میزان دقت را به چالش می‌کشد.

[1] Multilayer perceptron



نمودار (۱) خوشه‌ها (محور عمودی) و تعداد نگاشت‌ها بر تصاویر (محور افقی)



نمودار (۲) خوشه‌ها (محور عمودی) و تعداد نگاشت‌ها بر تصاویر (محور افقی)

```
[0.05       0.08947368 0.12894737 0.16842105 0.20789474 0.24736842
 0.28684211 0.32631579 0.36578947 0.40526316 0.44473684 0.48421053
 0.52368421 0.56315789 0.60263158 0.64210526 0.68157895 0.72105263
 0.76052632 0.8        ]
```

شکل (۱۲) دقت ماشین در محدوده ۰.۰۵ تا ۰.۸

سیستم بینش کمی در مورد حرکات اساسی دارد. رویکردهای پیچیده‌تر عبارتند از:

۱. تغذیه داده‌های ورودی در یک توالی از زمان از جمله داده‌های زمانی بیشتر
۲. استفاده از رویکرد CNN - LSTM
۳. استخراج ویژگی‌ها با CNN
۴. پس از آن گذر از این ویژگی‌ها
۵. تهیه نقشه‌هایی برای یادگیری و طبقه‌بندی دنباله‌های زمانی

رویکردهای کنونی در زمینه فعالیت‌های انسانی به دست آوردن حدود ۹۰ ٪ دقت در مجموعه داده‌های جمعیت می‌باشد.

در تجزیه و تحلیل طبقه‌بندی با مجموعه داده‌های خام در بخش اجرایی، ما نتایج آزمایش را با مجموعه داده‌های خام، یعنی سری زمانی گرفته‌شده از ماژول‌های مبتنی بر HIT برای طبقه بندی دسته‌های گره‌های دیتاست‌ها به کار می‌بریم.

یکی از روش‌های پیش بینی روش یادگیری مبتنی بر نمونه (الگوریتم k-نزدیکترین همسایه) و Bayes net در گذشته بهترین عملکرد طبقه بندی را نشان می‌داده‌اند. این نتیجه تعجب‌آور نیست، زیرا کارهای گذشته تشخیص جهت حرکت را با الگوریتم k-نزدیکترین همسایه انجام داده و دقت تشخیص خوبی را نشان می‌دهند. همچنین دستگاه بردار پشتیبانی، به ویژه با هسته خطی و پرسپترون چند لایه، عملکرد خوبی را نیز نشان می‌دهند. الگوریتم‌های دیگر به جز Bayes تقریباً بیش از ۹۸٪ دقت تشخیص دارند.

جدول (۴) مقایسه دقت تشخیص(٪)طبقه‌بندی خوشه بر اساس داده‌های خام

روش	۱۰۰۰ تایی اول	۱۰۰۰ تایی دوم	۱۰۰۰ تایی سوم	۱۰۰۰ تایی چهارم	۱۰۰۰ تایی پنجم
Bayes Net	79.51	78.58	75.43	80	79.41
CNN	78.19	77.18	77.89	77.80	76.03
LSTM	79.12	78.42	79.27	79.84	79.13
SVM	73.23	72.00	74.12	79.76	78.49
Our Proposed Method	80.12	80.14	71.01	81.56	82.01

جدول (**Error! No text of specified style in document.**)
مقایسه دقت تشخیص(٪) روش‌های مورد بررسی در تحقیق

روش‌های گوناگون	۱۰۰تایی اول	۱۰۰تایی دوم	۱۰۰تایی سوم	۱۰۰تایی چهارم
CNN	67.31	70.24	78.94	75.83
LSTM	72.49	75.29	76.08	75.35
Our Proposed Method	82.84	84.13	84.81	87.81

نتایج آزمایش را برای طبقه‌بندی خوشه‌های به‌دست آمده با توجه به تعداد پردازش دسته‌ها خلاصه می‌کنیم، یعنی نزدیک به ۱۰۰۰ تایی اول، نرسیده به هزار تایی دوم. در جدول فوق، می‌توانیم بدانیم که روش یادگیری مبتنی بر نمونه CNN و شبکه LTSM بهترین عملکرد طبقه بندی را نشان می‌دهند. سیستم پشتیبانی، به ویژه با دسته‌های درجه دوم و مکعب و پرسپترون چند لایه، نیز عملکرد خوبی نشان می‌دهد، اگرچه ممکن است این الگوریتم‌ها برای مجموعه داده‌های آموزش بار محاسباتی بیشتری نسبت به‌الگوریتم k-نزدیکترین همسایه داشته باشند.

از نتایج تجربی، میتوانیم بدانیم که هر ۱۰۰۰ تایی مبتنی بر HIT که بر روی دیواره‌های مقابل یکدیگر قرار گرفته‌اند، می‌توانند جهت حرکت، فاصله بدن از سنسورها و سرعت حرکت را در دو طرف دسته‌بندی کنند، یک ۱۰۰۰ تایی مبتنی بر HIT که می‌تواند جهت، مسافت و سرعت افراد را در حال راه رفتن طبقه‌بندی کند، اما شناسایی موضوع را به خوبی انجام نمی‌دهد. بر این اساس، ما می‌توانیم تست‌های این مطالعه را که برای ایجاد یک محیط هوشمند سازگار است، تصور کنیم که در آن مجموعه‌ای از HITهای در ترکیب با سوالات به هم متصل شده‌اند و چندین ماژول مبتنی بر HIT در یک مجموعه توزیع شده‌است.

نتیجه گیری

می‌بینیم که با یک شبکه CNN، می‌توانیم به دقت طبقه‌بندی ۷۵ ٪ روی داده دست یابیم. مجموعه داده ۵ کلاس نشان‌دهنده اثربخشی شبکه‌های عصبی بر روی مشکلات دیتاست است. حتی اگر اجرای ما فقط ورودی‌های سنسور را در زمان اجرای خاص تبلیغ کند و با توجه به این واقعیت که دنباله‌ها ممکن است به دقت معقول دست یافته باشد. ویژگی‌های موقتی در داده‌های حسگر قرار داده می‌شود و بدین ترتیب امکان بدست آوردن مدل را فراهم می‌آورد. ما می‌دانیم که افراد غیر حرفه‌ای واقعاً در پیش بینی رفتارهای انسانی فقط با استفاده از چند نمونه "آموزش" ممکن است بهتر عمل کنند، در حالی که کارهای قبلی به طور معمول تفسیرپذیری یا نقش آن‌ها را در بهبود اعتماد انسان ارزیابی کرده‌اند، فرضیه اولیه ما این است که کارهای زیادی برای توسعه روش‌های بهبود یافته برای بهبود توانایی در تیم‌های هوش مصنوعی-انسان باقی مانده‌است. کارهای آینده می‌تواند شامل نا محدود کردن خوشه‌ها و ارزیابی میزان بهبود عملکرد انسانی در FP و KP به موفقیت بهتر دسته‌های AI-human در تحقق یک هدف مشترک کمک کند. از منظر کارایی، سیستم‌های VQA به دلیل ارتباط مشابه کارهای روزمره انسانی، آینده بسیار روشنی دارند. این سیستم می‌توان هزینه و شرایط پیچیده بسیاری از مشاغل و فعالیت‌ها را بهبود بخشد و به عنوان جزئی حیاتی در زندگی بشر مورد استفاده قرار گیرد.

مراجع

[1] M. Cristani, R. Raghavendra, A. Del Bue and V. Murino, "Human Behavior Analysis in Video Surveillance: A Social Signal Processing Perspective," *ScienceDirect, Neurocomputing,* vol. 100, pp. 86-97, 2012.

[2] S. Sreela and M. Sumam, "Action Recognition in Still Images using Residual Neural Network Features," *International Conference on Advances in Computing and Communication Procedia Computer Science in Science Direct,* vol. 143, pp. 563-569, 2018.

[3] Y. Quan, R. Chen, R. Xu and H. Ji, "Attention with structure regularization for action recognition," *Science Direct Computer Vision and Image Understanding,* vol. 102794, 2019.

[4] M. Masala, S. Ruseti and T. Rebedea, "Sentence selection with neural networks using string kernels string kernels," *Science Direct Procedia Computer Science,* vol. 112, pp. 1774-1782, 2017.

[5] A. Das, H. Agrawal, L. Zitnick, D. Parikh and D. Batra, "Human Attention in Visual Question Answering: Do Humans and Deep Networks Look at the Same Regions," *Science Direct Computer Vision and Image Understanding,* 2016.

[6] A. Singh, V. Natarajan, M. Shah, Y. Jiang, X. Chen, D. Batra, D. Parikh and M. Rohrbach, "Towards VQA Models That Can Read," *In CVPR,* 2019.

[7] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick and D. Parikh, "Visual Question Answering," *International Conference on Computer Vision (ICCV),* 2015.

[8] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam and M. Lanctot, "Mastering the game of Go with deep neural networks and tree search," *Nature,* vol. 529, pp. 484-489, 2016.

[9] "OpenAI ve," 2018. [Online]. Available: https://blog.openai.com/openai-five/.

[10] E. Kolve, R. Mottaghi, D. Gordon, Y. Zhu, A. Gupta and A. Farhadi, "Ai2-thor: An interactive 3d environment for visual ai," *ArXiv,* Vols. 1712-05474, 2017.

[11] A. Das, S. Datta, G. Gkioxari, S. Lee, D. Parikh and D. Batra, "Embodied Question Answering," *IEEE Conference on Computer Vision and Pattern Recognition,* 2018.

[12] Y. Duan, X. Chen, R. Houthooft, J. Schulman and P. Abbeel, "Benchmarking deep reinforcement learning for continuous control," *CoRR,* Vols. abs-1604.06778, 2016.

[13] G. Brockman, V. Cheung, L. Pettersson, J. Schneider, J. Schulman, J. Tang and W. Zaremba, "Openai gym," *CoRR,* Vols. abs-1606,01540, 2016.

[14] L. Yu, E. Park, A. Berg and T. Berg, "Visual madlibs: Fill-in-the-blank description generation and question answering," in *Computer Vision (ICCV)*, 2015.

[15] P. Zhang, Y. Goyal, D. Summers-Stay, D. Batra, D. Parikh and Y. Yang, "Balancing and answering binary visual questions," *Computer Vision and Pattern Recognition (CVPR),* vol. abs/1511.05099, 2015.

[16] C. Zitnick and D. Parikh, "Bringing Semantics Into Focus Using Visual Abstraction," *Computer Vision and Pattern Recognition (CVPR),* 2013.

[17] C. Zitnick, D. Parikh and L. Vanderwende, "Learning the Visual Interpretation of Sentences," *Computer Vision (ICCV),* pp. 1681-1688, 2013.

[18] C. Zitnick, R. Vedantam and D. Parikh, "Adopting Abstract Images for Semantic Scene Understanding," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 38, no. 4, pp. 627 - 638, 2016.

[19] X. He, Y. Zhang, L. Mou, E. Xing and P. Xie, "PathVQA: 30000+ Questions for Medical Visual Question Answering," *arXiv,* vol. 2003.10286, 2020.