بسمه تعالی

بهبود پاسخ دهی به متون درون عکس با تصحیح توکن‌های استخراج شده از تصویر

کبری فرشیدی*، حسن ختن لو، محرم منصوری‌زاده

گروه مهندسی کامپیوتر، دانشکده مهندسی، دانشگاه بوعلی سینا، همدان

k.farshidi@eng.basu.ac.ir

خلاصه:

یکی از انواع مسائل چند وجهی، مسئله‌ی پاسخ‌گویی به متون درون تصویر است (شکل ۱) . هدف، کشف ارتباط بین تصویر (بخصوص متون و علائم درون تصویر) و سوال متنی پرسیده شده، درباره‌ی آن تصویر می باشد. در نتیجه‌ی این پژوهش ساخت دستیار هوشمند برای معلولان، نابینایان و کودکان خردسال است حتی از ثمره ی این پژوهش میتوان در ماشین‌های تسلا بهره برد. پژوهش‌گران توسط مدل‌های یادگیری عمیق، معماری مبدل‌ها، مکانیزم توجه و ... توانسته‌اند به نتایج خوبی دست پیدا کنند. اما هنوز این حوزه بصورت عام و گسترده بکار برده نشده است. چرا که دقت به حد قابل قبولی نرسیده است. یکی از دلایل دقت پایین، ورود توکن های استخراج شده ی اشتباه در مدل میباشد. چرا که موتور استخراج‌کننده‌ی توکن هر چقدر هم قوی بوده باشد، باز هم دارای خطا در تولید توکن ها است. در این پژوهش قصد داریم با به‌کار گیری پیش پردازشها و تصحیح توکن‌های ورودی این دقت را بالا ببریم.

کلمات کلیدی:

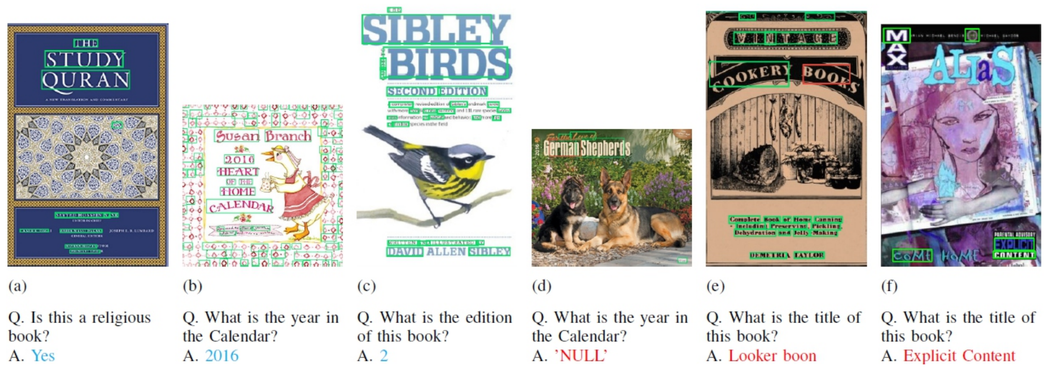
پرسش از متون درون تصویر، مبدل، مکانیزم توجه، مدل از پیش‌آموزش دیده

مقدمه:

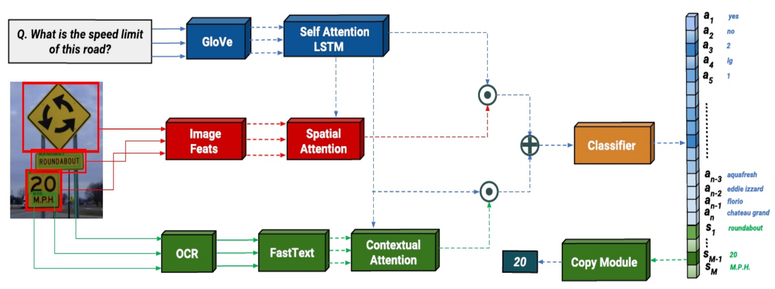
از سال ۲۰۱۹ تا کنون پژوهش‌های متفاوتی در زمینه‌ی متون درون تصویر انجام شده است. تلاشهای پژوهش‌گران در ۴ سال اخیر را میتوان به ۳ دسته تقسیم کرد . دسته ی اول بهره گیری از موتور تشخیص کاراکتر و ادغام آن توسط مکانیزم توجه با ویژگی های متن و تصویر میباشد (شکل ۲) دسته ی بعد علاوه بر بکارگیری موتور ocr از معماری مبدلها که درسال ۲۰۲۰ از معماری نوظهور آن زمان بود، بهره بردند و دقت این حوزه را به حد قابل قبولی رساندند (شکل ۳) .دسته سوم از تکنولوژی‌های پیش‌آموزش بهره بردند تا بامدل‌هایی با دقت بالاتر معماری خود را غنی کنند(مدل‌هایی که با داده های بسیار زیاد از قبل آموزش دیده شده‌اند.)(شکل ۴)

از بعد دیگر پژوهش‌گران تلاش کردند تا از همان ابتدا مجموعه داده های خوب و معیار های ارزیابی متناسب ایجاد کنند. .

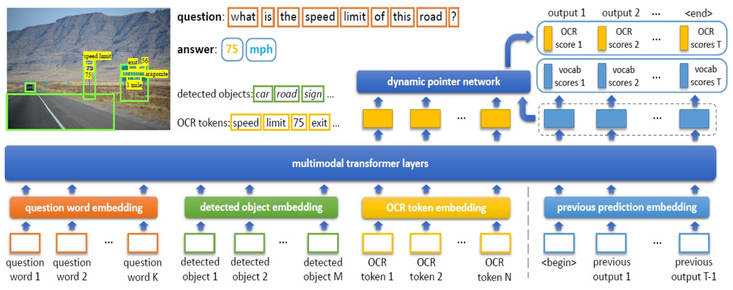
آنچه قابل توجه است تکنولوژی روبه روز به سمت مدل‌های پیش آموزش جلو میرود تا ویژگی‌های استخراج شده با دقت بالاتری بدست آید لذا ما نیز از دسته سوم بهره میبریم با این تفاوت که در حین آموزش، دقت ویژگی‌های توکن‌های استخراج شده را بالا میبریم.



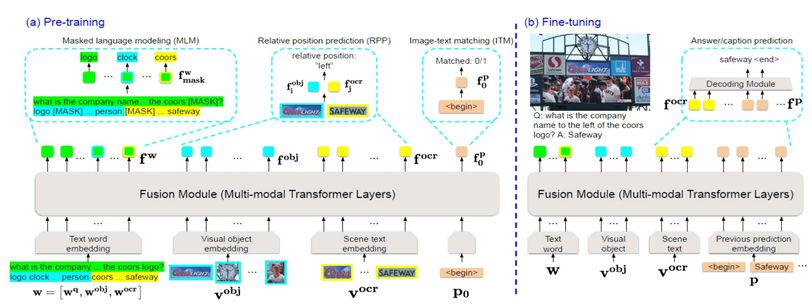
شکل ۱:



شکل ۲:



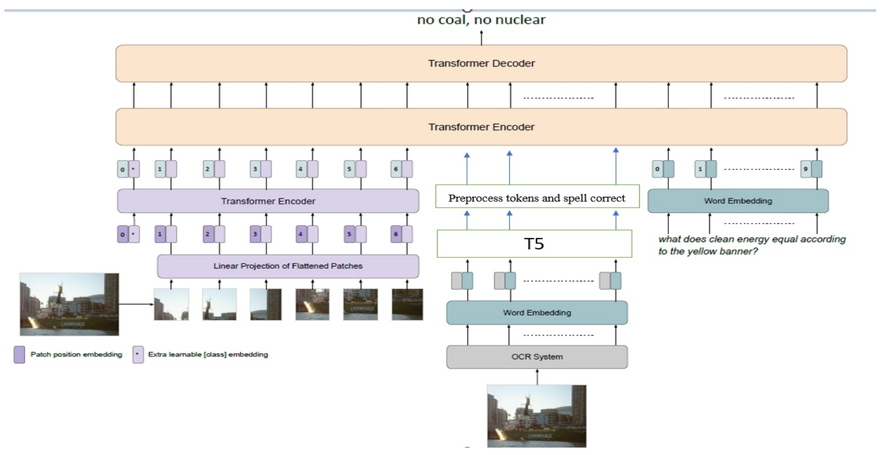
شکل ۳:



شکل ۴:

مواد و روشها:

همانطور که گفته شد ما از دسته سوم پژوهشگران بهره میبریم و دقت آن رابالاتر می‌بریم. همانند شکل روبرو (شکل ۵) به این صورت عمل می کنیم که قبل از اینکه ویژگی‌های استخراج شده وارد معماری مبدل شود و همه ی وجه ها با هم ادغام شوند ، ویژگی‌های توکن‌ها را بهبود می‌دهیم چرا که در صورتی که توکن‌های اشتباه وارد معماری مبدل شود قطعا جواب‌های اشتباهی نیز استخراج میشود. و استدلال‌های اشتباهی از متون درون تصویر میشود. مدل‌های پیش آموزش دیده‌ی متعددی به جهت بررسی صحت توکن ها و جملات بوجود آمده است که میتوان با استفاده از آنها توکن‌ها را بهبود داد و از طرفی با یک پیش پردازش متون استخراج شده از تصویر جملات بهتری برای استدلال بوجود آورد.



شکل ۵:

بحث و نتیجه‌گیری:

چالش اساسی پژوهش‌گران بیشتر آماده‌سازی مجموعه داده‌های مرتبط با این پژوهش ، مجموعه داده ساده‌تر و در آن واحد دارای اطلاعات غنی‌تر، طراحی معماری‌های جدیدتر با توجه به پیشرفت معماری‌های شبکه‌های عمیق برای بالابردن عملکرد این حوزه میباشد. ما درمعماری خود سعی کرده ایم با بهره‌گیری بهتر از توکن‌های استخراج شده از طریق موتور جستجوگر، دقت این حوزه را بالا ببریم.

لیست اختصارات:

Visual Question Answering=VQA
Text Based Visual Question Answering=TVQA
Optical Character Recognition=OCR

منابع و مراجع:

[1] amanpreet S., Vivek N., Meet S., Yu J., Xinlei C., Dhruv B., Devi P. and Marcus R., Towards VQA model that can read. IEEE 2020

[2] Ronghang Hu., Amanpreet S., Trevor D,. Iterative answer prediction with poiter-argumented multimodal transformer or TVQA.IEEE 2020

[3] Chenyu G.,Qi Z., PengWang, Hui L., Yuliang L., Anton V. . Structured mutlimodal attention for TVQA.arXiv.2020

[4] Zhengyuan Y., Yijuan L., Jianfeng W.,Xi Y., Dinei F., Lijuan W., Cha Z., Lei Z., TAP: Text-aware pretraining for TVQA and Text caption. arXiv 2020

[5] Furkan A., Litman R., Xie Y., appalarajau S., . LaTr: Layout-aware Tranformer for Scene-Text VQA.arXiv.2021