دانشگاه صنعتی امیرکبیر
(پلی تکنیک تهران)

دانشکده مهندسی کامپیوتر و فناوری اطلاعات

پرسش و پاسخ بصری با استفاده از شبکه های عصبی کانولوشنی و بازگشتی عمیق

علی غلامی

۱۶ اردیبهشت ۱۳۹۷

۱ مقدمه

درك انسان از محيط اطراف خود، از ابتداي كودكی شكل می گيرد. اگر از يک كودك، درباره جهانی كه براي او قابل رويت است، سوالی پرسيده شود، بلافاصله درباره ي اين جهان جملاتی توصيفی با دقت بسيار بالا ارائه می كند. انتقال اين قدرت به ماشين ها از اين جهات مختلفی حائز اهميت می باشد. يكی از جهات اصلی آن، كاربرد هاي بی شماري است كه می توان از اين قابليت در امور مختلفی استفاده كرد. دوربين هاي نظارتی، خودروهاي خودران و ... نمونه هايی از اين كاربرد ها هستند. درك تصاوير به همراه درك سوالاتی كه مرتبط با آن تصاوير پرسيده می شود، می تواند يكی از اساسی ترين قابليت هاي يک سامانه مديريت هوشمند تصاوير، خودروي خودران و يا يک سيستم بازيابی تصوير هوشمند باشد. براي اين منظور نياز است كه ابتدا صحنه به نمايش درآمده توسط ماشين هضم گردد. بدين معنی كه بدون درك درست از آنچه در تصوير موجود است، ساخت چنين سامانه اي امكان پذير نخواهد بود

۲ مرور سوابق پيشين

روش هايی كه جهت حل مسائل پرسش و پاسخ بصري ارائه شده اند از تنوع بالايی برخوردار هستند. اگرچه، می توان اغلب اين روش ها را در راستاي حل چهار چالش مهم زير در نظر گرفت:

۱. استخراج ويژگی از تصاوير
۲. استخراج ويژگی از متن و سوالات پرسيده شده
۳. تركيب بردار هاي ويژگی مستخرج از تصوير و متن
۴. چالش توليد پاسخ متناسب با ويژگی ها تركيب شده

با تحول عظيمی كه در سال ۲۰۱۲ با ارائه مدل كانولوشنی عميق جهت استخراج ويژگی ارائه شد و نيز افزايش قدرت پردازشی ماشين ها، توجه به سمت شبكه هاي كانولوشنی جهت استخراج ويژگی از تصاوير بيشتر شد. با پيشرفت اين مدل ها و تركيب آنها با مدل هاي پردازش زبان طبيعی، توليد شرح بر تصاوير نيز مورد توجه قرار گرفت. از سال ۲۰۱۵ تا به

اکنون، یکی از جالب ترین موضوعاتی که توجه پژوهشگران را به سمت خود جلب کرده است، موضوع پرسش و پاسخ بصري می باشد. این موضوع ارتباط تنگاتنگی با موضوع تولید شرح بر تصاویر دارد. به همین منظور، بسیاري از تکنییک هاي رایج در بحث شرح بر تصاویر، در موضوع پرسش و پاسخ بصري نیز مورد توجه قرار گرفته است.

۳ طرح پیشنهادی

سیستم پرسش و پاسخ بصري مطرح در این گزارش، به زبان پایتون و با استفاده از چارچوب کاري تنسورفلو پیاده سازي خواهد شد. هسته این چارچوب کاري به زبان سی پلاس پلاس و با استفاده از پلتفرم توسعه موازي کودا پیاده سازي شده است. این چارچوب کاري توسط تیم گوگل برین در حال توسعه بوده و پشتیبانی می شود. ایده ي اصلی مطرح در این چارچوب کاري، بیان محاسبات در قالب گراف است. هر گره ي این گراف، یک واحد محاسباتی را مشخص می کند. با این رویکرد می توان شبکه هاي پیچیده را به راحتی پیاده سازي نمود.

۴ محصولات طرح

خروجی هاي این طرح شامل موارد زیر خواهند بود:

- بازیابی تصاویر در شبکه های اجتماعی با استفاده از جستجوی متن
- توصیف کننده جهان اطراف جهت استفاده نا بینایان و مذاکره دو طرفه
- ربات های امداد گر در عملیات هایی که با انسان تبادل اطلاعات انجام می دهند
- ربات های مذاکره کننده

۵ مراحل انجام

۱. شناسایی و تهیه منابع:

- پیدا کردن مقالات مربوط به پرسش و پاسخ بصری
- یافتن توضیحات مربوط به طرح های از پیش پیاده سازی شده

۲. تنظیم ساختار:

- تهیه ي بخشهاي مقدمه، محتواي اصلی، چیکده، نتیجه گیري و منابع
- مشخص کردن ترتیب مباحث محتوا
- تهیه ي فهرست مباحث اصلی و فرعی یافتن معماری شبکه های کانولوشنی

۳. مطالعه و یادداشت برداری:

- پیدا کردن مقالات مربوط به پرسش و پاسخ بصری
- یافتن توضیحات مربوط به طرح های از پیش پیاده سازی شده

۴. اجرای بخش عملی:

- پیاده سازی شبکه های کانولوشنی عمیق
- پیاده سازی شبکه بازگشتی عمیق
- آموزش شبکه
- آزمایش و ارزیابی

۵. تهیه گزارش نهایی:

- مستند سازي و نوشتن گزارش نهایی
- آمادگی جهت ارائه ي شفاهی

۶ زمان بندی پروژه

زمان بندی پروژه بر اساس ”پیشنهاد زمان بندی پروژه کارشناسی“ در کتاب انجام شده است.