پاسخ به پرسش دیداری در تصاویر هنری با استفاده از یادگیری عمیق

عرفان ذوالقدری‌ها[۱]، کاظم فولادی قلعه[۲] و پویا اردەخانی[۳]

[۱]دانشجوی کارشناسی ارشد، آزمایشگاه پژوهشی یادگیری عمیق،
گروه مهندسی کامپیوتر، دانشکده مهندسی، دانشکدگان فارابی، دانشگاه تهران
erfanzolghadriha@gmail.com

[۲] استادیار، گروه مهندسی کامپیوتر، دانشکده مهندسی، دانشکدگان فارابی، دانشگاه تهران
kfouladi@ut.ac.ir

[۳]دانشجوی کارشناسی، آزمایشگاه پژوهشی یادگیری عمیق،
گروه مهندسی کامپیوتر، دانشکده مهندسی، دانشکدگان فارابی، دانشگاه تهران
pouya.ardehkhani@ut.ac.ir

چکیده‌ـ پاسخ به پرسش دیداری در حوزه‌های خاص علاوه بر تازگی، از این رو که به کاربردی‌تر شدن این سیستم‌ها در مسائل روزمره و مسائل تخصصی کمک می‌کند، اهمیت دارد. در این پژوهش با استفاده از یک مجموعه‌داده هنری که دارای پرسش‌های دیداری و برمبنای دانش می‌باشد، اقدام به پیاده‌سازی و بهبود عملکرد یک سیستم پاسخ به پرسش دیداری در تصاویر هنری می‌کنیم. برای این کار در ابتدا ماهیت پرسش‌های مجموعه‌داده را با استفاده از یک BERT پیش آموزش دیده مشخص کرده و سپس در شاخه‌ی دیداری با استفاده از مدل iQAN با مکانیسم توجه MLB و مکانیسم همجوشی MUTAN به پرسش‌های دیداری و در شاخه‌ی مبتنی بر دانش با استفاده از یک مدل مبتنی بر XLNet به پرسش‌هایی که از روی تصاویر قادر به پاسخ‌دهی به آن‌ها نیستیم، پاسخ می‌دهیم. در شاخه‌ی دیداری به دقت ۷۸/۹۲ درصد در پرسش‌های دیداری رسیدیم. در شاخه‌ی مبتنی بر دانش نیز به دقت ۴۷/۷۱ درصد دست پیدا کردیم. در مجموع دو شاخه با توجه به تقسیم آزمایشی مجموعه‌داده به دقت ۵۵/۸۸ درصد رسیدیم. همچنین در این پژوهش تأثیر پارامترهای تعداد نگاه اجمالی و توابع فعال‌سازی را در عملکرد مدل بررسی شده است.

کلمات کلیدی– بینایی کامپیوتری، پاسخ به پرسش دیداری، پردازش زبان طبیعی، تصاویر هنری

۱. مقدمه

پاسخ به پرسش دیداری یک مفهوم نسبتاً جدید و مهیج در مبحث یادگیری عمیق می‌باشد. پاسخ به پرسش دیداری را می‌توان به عنوان بسط مفهوم درک ماشین در نظر گرفت. همچنین پاسخ به پرسش دیداری یک وظیفه‌ی هوش مصنوعی چند زمینه‌ای است که تکنیک‌های پیشرفته بینایی کامپیوتری و پردازش زبان طبیعی را برای ساختن سیستمی ترکیب می‌کند که می‌تواند به پرسشی درباره یک تصویر پاسخ دهد. این مدل تلاش می‌کند تا معنا و معناشناسی پشت تصویر را درک کند و بر اساس «درک» آن به پرسش‌ها پاسخ دهد. اکثر تحقیقات گسترده‌ای که در این زمینه انجام می‌شود، باعث می‌شود از معماری‌های یادگیری عمیق و الگوریتم‌های یادگیری مختلف در حوزه‌های بینایی کامپوتری، تشخیص اشیا و پردازش زبان طبیعی برای دستیابی به هدف استفاده شود. اغلب مطالعات انجام شده در حوزه پاسخ به پرسش دیداری به مجموعه‌داده‌هایی اعمال شده که از نظر محتوایی عمدتاً دارای دسته بندی خاصی نیستند. به این صورت که تصاویر در دسته بندی‌های مختلفی از قبیل اتومبیل‌ها، اشخاص، مکان‌ها، حیوانات و... در درون یک مجموعه‌داده قرار می‌گیرند. از همین رو، اکثر مطالعات و تحقیقات صورت گرفته در این حوزه به ارائه الگوریتمی پرداخته‌اند که بتواند به پرسش‌هایی پاسخ دهد که با اطلاعاتی که از تصاویر کسب می‌شوند قادر به پاسخ‌دهی باشد. اخیراً توجه به پاسخ‌دهی به پرسش دیداری در حوزه‌های خاص‌تر از قبیل

پزشکی، هنری و... که نیاز به پردازش تصاویر و پرسش‌های محدود به آن زمینه را دارد، رو به افزایش است. این توجه باعث افزایش کاربردی شدن این سیستم‌ها در زمینه‌هایی که برای پاسخ به پرسش‌های تصویری نیازمند داشتن دانش در آن زمینه است، می‌شود. موضوع هنر و بینایی کامپیوتری نیز از این رو که بسیاری از عناصر هنری دارای اجزای دیداری می‌باشند، پیوندی اجتناب ناپذیر دارند. دیجیتالی کردن آثار هنری برای نگهداری و ترمیم آن‌ها، یک قدم اساسی در این حوزه می‌باشد. تا کنون مطالعات متعددی در حوزه‌ی بینایی کامپیوتری بر روی آثار هنری انجام شده است. این‌کارها شامل وظایف شناسایی سبک و صاحب اثر [1,2]، دسته بندی تصاویر [3,4,5,6,7] و بازیابی تصاویر [8,9,10] می‌شوند. در سال ۲۰۱۸، مجموعه‌داده SemArt در [11] معرفی شد. این مجموعه‌ی داده در اصل برای درک معنایی هنر ارائه شده است و شامل نقاشی‌ها و نظرات مربوط به آن‌ها می‌باشد. این نظرات بلوک‌های از متن هستند که شامل ابرداده‌های مربوط به نقاشی می‌باشند. در سال ۲۰۲۰ در [12] مجموعه‌داده‌ی AQUA برای وظیفه‌ی پاسخ‌دهی به پرسش دیداری در تصاویر هنری از روی مجموعه‌داده‌ی SemArt ساخته شد. از آن جایی که برای ایجاد پرسش‌ها در این مجموعه‌داده از تکنیک‌هایی استفاده شده که نه تنها بر روی محتوای دیداری در خود نقاشی، بلکه بر روی نظرات مربوط به آن‌ها نیز تمرکز می‌کنند، جفت‌های پرسش و پاسخ موجود در این مجموعه‌داده دارای حالات دیداری و نیازمند به دانش خارج از تصاویر هستند. در همان مقاله [12] که مجموعه‌داده‌ی AQUA معرفی شد، یک مدل پایه (VIKING) نیز برای پاسخ‌دهی به پرسش‌های این مجموعه‌داده ارائه شد که این کار، اولین تلاش صورت گرفته برای حل مسئله‌ی پاسخ به پرسش دیداری در تصاویر هنری می‌باشد. مدل VIKING از سه بخش کلی ایجاد شده است. در بخش که بخش انتخاب ماهیت نام دارد، ماهیت پرسش‌ها مشخص شده و پرسش‌ها به دو دسته‌ی دیداری و نیازمند به دانش خارجی تبدیل می‌شوند. در شاخه‌ی دیداری برای پرسش‌هایی که از روی تصاویر قادر به پاسخ‌گویی آن‌ها هستیم با استفاده از مدل iQAN پایه [13]، پاسخی پیش‌بینی می‌شود. iQAN یک مدل دوگانه است و می‌تواند یک سؤال یا یک پاسخ را به عنوان ورودی بگیرد، سپس همتای آن را به عنوان خروجی ایجاد کند. بخشی از پرسش‌های دیداری مجموعه‌داده AQUA نیز با استفاده از این مدل تولیده شده‌اند. پرسش‌های طبقه‌بندی شده به عنوان دانش محور به شاخه پاسخ به پرسش برمبنای دانش خارجی داده می‌شوند. این شاخه دارای دو بخش می‌باشد. در بخش اول یک بازیابی دانش خارجی دو مرحله‌ای صورت می‌گیرد که نظرات مربوط به هر پرسش یافته شده و بنابر ارتباط آن‌ها با پرسش مربوطه رتبه‌بندی می‌شوند، و در نهایت زیرمجموعه‌ای از ده نظر مرتبط برای هر پرسش ایجاد می‌شود. در بخش بعدی نیز برای پرسش‌های مبتنی بر دانش پاسخی با استفاده از یک مدل XLNet [14] پیش‌بینی می‌شود. در این مقاله با استفاده از مجموعه‌داده AQUA اقدام به پیاده‌سازی و بهبود عملکرد مدل VIKING می‌کنیم. برای این کار ما مکانیسم توجه[1] MUTAN [15] مورد استفاده در مدل پایه را با مکانیسم توجه MLB [16] جایگزین می‌کنیم و از آن در کنار مکانیسم همجوشی[2] MUTAN در مدل iQAN بهره می‌بریم. کدهای مدل پایه و مجموعه‌داده AQUA در دسترس[3] هستند. در بخش ۲ به طور جامع‌تر به معرفی روش پیشنهادی می‌پردازیم.

۲. روش پیشنهادی

در این کار از مدل پایه‌ی ارائه شده در [12] برای حل مسئله پاسخ به پرسش دیداری در تصاویر هنری و اعمال برخی تغییرات برای ایجاد بهبود در عملکرد این مدل استفاده شده است. این مدل سه بخش کلی دارد؛ بخش انتخاب ماهیت، شاخه‌ی دیداری و شاخه‌ی مبتنی بر دانش که در ادامه به معرفی آن‌ها می‌پردازیم.

۲-۱- *انتخاب ماهیت*

در این بخش به معرفی روش مورد استفاده در شاخه انتخاب ماهیت برای رمزگذاری و استخراج بردارهای ویژگی از پرسش‌ها و نقاشی‌های مجموعه‌داده و در نهایت الحاق این ویژگی‌ها برای انتخاب ماهیت پرسش‌ها می‌پردازیم. در این قسمت برای استخراج ویژگی‌های متنی از یک مدل BERT پیش آموزش دیده [17] به عنوان رمزگذار پرسش‌ها استفاده می‌شود. مدل انتخابی در مدل پایه BERT-Large, Uncased می‌باشد که در نهایت با استفاده از آن پرسش را به یک بردار ۱۰۲۴ بعدی q کدگذاری می‌کنیم. برای استخراج ویژگی‌های نقاشی‌ها نیز از معماری ResNet-152 پیش آموزش دیده [18] استفاده می‌کنیم تا نقاشی‌ها را به یک بردار ۲۰۴۸ بعدی v کدگذاری کنیم. سپس با استفاده از یک انتخابگر ماهیت S یک پرسش q را با استفاده از بردارهای v و q به یکی از دو دسته‌ی پرسش‌های دیداری و یا پرسش‌هایی که برای پاسخ‌دهی به آن‌ها نیازمند دانش خارجی هستیم تقسیم می‌کنیم. برای این کار بردارهای ویژگی v و q را به یک دیگر الحاق کرده و بردار X را ایجاد می‌کنیم و با استفاده از این بردار در یک مدل رگرسیون لجستیک[4] به صورت فرمول (۱) عملیات انتخاب ماهیت را انجام می‌دهیم.

[1] Attention
[2] Fusion
[3] https://github.com/noagarcia/ArtVQA
[4] Logistic regression

$$S(X) = \frac{1}{1+e^{-(w_S^T X - b_S)}} \quad (۱)$$

در فرمول (۱) پارامترهای w_s و b_s بردارهای عددی آموزش پذیر هستند. یک پرسش q در صورتی به شاخه‌ی پاسخ به پرسش دیداری داده می‌شود که در آن مقدار $S(X) > 0.5$ باشد. در غیر این صورت پرسش q به شاخه‌ی برمبنای دانش ارائه می‌شود.

۲-۲- *شاخه پاسخ به پرسش دیداری*

پرسش‌های دیداری را می‌توان تنها بر اساس نقاشی مرتبط به آن‌ها، بدون هیچ دانش خارجی پاسخ داد. برای این نوع پرسش‌ها، کار به پاسخ به پرسش دیداری بر روی نقاشی‌ها کاهش می‌یابد. در این پژوهش در مدل iQAN از ماژول مکانیسم توجه MLB به جای ماژول مکانیسم توجه MUTAN، به همراه مکانیسم همجوشی MUTAN به عنوان شاخه پاسخ به پرسش دیداری خود استفاده می‌کنیم. ماژول توجه وظیفه‌ی افزایش تمرکز بر روی نقاطی از تصاویر ورودی که به پرسش مرتبط‌تر هستند را دارد. ماژول همجوشی نیز وظیفه تلفیق ویژگی‌های متنی و دیداری را برای رسیدن به بردارهای پاسخ بر عهده دارد. برای این کار به طور جداگانه مدل iQAN را بر روی تقسیم آموزشی مجموعه‌داده AQUA آموزش می‌دهیم. این شاخه یک پاسخ پیش‌بینی شده a_v را تولید می‌کند که از واژگان پاسخ A متشکل از ۵۰۰۰ کلمه رایج در تقسیم آموزشی می‌باشد.

مدل iQAN یا شبکه معکوس پذیر پاسخ‌دهی به پرسش یک مدل سرتاسر یکپارچه می‌باشد. در این مدل با استفاده از ماژول همجوشی دو خطی معکوس پذیر و طرح به اشتراک گذاری پارامتر، می‌توان هم وظیفه‌ی پاسخ به پرسش دیداری و هم وظیفه تولید پرسش دیداری را به طور همزمان انجام داد. با آموزش مشترک این دو وظیفه با تنظیم‌کننده‌های دوگانه (آموزش دوگانه)، این مدل درک بهتری از تعاملات بین تصاویر، پرسش‌ها و پاسخ‌ها خواهد داشت. پس از آموزش، iQAN می‌تواند پرسش یا پاسخ را به عنوان ورودی دریافت کند و همتای آن را در خروجی ایجاد کند.

در این پژوهش در قسمت پاسخ‌دهی به پرسش دیداری، با توجه به یک پرسش، یک RNN برای به دست آوردن ویژگی تعبیه شده q، و در مقابل از CNN برای تبدیل تصویر ورودی به یک نقشه ویژگی استفاده می‌شود. یک ماژول توجه مبتنی بر مدل MLB [17] برای ایجاد یک بردار ویژگی تصویری آگاه از پرسش v_q مورد استفاده قرار می‌گیرد. سپس با استفاده از یک ماژول همجوشی MUTAN [18] دیگر بردارهای ویژگی پاسخ $\hat{a}$ با ادغام v_q و q به دست می‌آیند. در نهایت، یک طبقه‌بندی کننده خطی W_a پاسخ را پیش‌بینی می‌کند. قسمت پاسخ به پرسش دیداری در مدل پایه‌ی iQAN بر اساس یکی از مدل‌های پیشرفته‌ی پاسخ به پرسش دیداری یعنی MUTAN است. ما در این پژوهش از مکانیستم توجه پیاده‌سازی شده‌ی مدل پاسخ به پرسش دیداری MLB در ترکیب با مکانیسم همجوشی MUTAN استفاده کردیم و از این جهت، مدل iQAN را دستخوش تغییر قرار دادیم.

۳-۲- *شاخه پاسخ به پرسش بر مبنای دانش خارجی*

پرسش‌های طبقه‌بندی شده به عنوان دانش محور به شاخه پاسخ به پرسش برمبنای دانش خارجی داده می‌شوند. در این شاخه در ابتدا نظری که در C بیشترین ارتباط را با q دارد با یک استراتژی دو مرحله‌ای بازیابی می‌شود. در مرحله اول، از TF-IDF برای رتبه‌بندی همه نظرات در C با توجه به ارتباط آن‌ها با q در مدل پایه استفاده می‌شود و زیرمجموعه C_q را شامل ۱۰ نظر مرتبط برتر بین همه c_iهای عضو C به دست می‌آید. در این بخش، اگر $\hat{q}$ و $\hat{c}_i$ بردارهای TF-IDF مربوطه باشند، امتیاز را مطابق فرمول (۲) به صورت زیر محاسبه می‌کنیم:

$$S_i = \hat{q}^T \hat{c}_i / (||\hat{q}|| \, ||\hat{c}_i||) \quad (۲)$$

مجموعه C_q متشکل از ۱۰ c_i است که بالاترین S_i را دارند. به منظور بهبود دقت رتبه‌بندی در پیش‌پردازش q و c_i از NLTK برای توقف حذف کلمه و ریشه‌یابی کلمه و از TF-IDF در حالت n-grams که در آن n = 3 می‌باشد، استفاده می‌شود.

مرحله دوم، برای یافتن نظر مرتبط با پرسش، c_i عضو C را رتبه‌بندی می‌کند. این مسئله به عنوان یک مسئله طبقه‌بندی دو جمله در نظر گرفته می‌شود و از یک مدل مبتنی بر BERT (bert-base-uncased) برای پیش‌بینی احتمال اینکه چقدر یک c_i داده شده با q مرتبط است، استفاده می‌شود. q و هر c_i عضو C با استفاده از [SEP] به یک دیگر الحاق شده و سپس به BERT از پیش آموزش دیده اعمال می‌شوند. BERT از نشانه‌های ویژه [CLS] و [SEP] برای درک صحیح ورودی استفاده می‌کند. خروجی o_i مرتبط با [CLS] به یک مدل رگرسیون لجستیک ارسال می شود که توسط فرمول (۳) به دست می‌آید:

$$R(o_i) = \frac{1}{1+e^{-(w_r^T o_i - b_r)}} \quad (۳)$$

در فرمول (۳) w_r و b_r بردارهای آموزش پذیر و عددی هستند. این مدل به عنوان یک طبقه‌بندی‌کننده باینری آموزش داده می‌شود که پیش‌بینی می‌کند آیا q و c_i مرتبط هستند یا خیر، اما خروجی $R(o_i)$ آن به عنوان امتیاز c_i در هنگام استنتاج در نظر گرفته می‌شود. در مدل پایه از c_{i^*} زمانی که $i^* = \arg\max_i R(o_i)$ می‌باشد برای پاسخ دادن به پرسش استفاده می‌شود.

در مدل پایه، برای پیش‌بینی پاسخ پرسش‌های مبتنی بر دانش خارجی از XLNet [14] استفاده می‌شود. XLNet یک ترانسفورمر

اتورگرسیو است و توسعه‌ای از مدل Transformer-XL [19] می‌باشد که از بهترین مدل‌سازی زبان اتورگرسیو[5] و رمزگذاری خودکار استفاده می‌کند و در عین حال تلاش می‌کند از محدودیت‌های آن‌ها اجتناب کند. این مدل از قبل با استفاده از روش خودبازگشتی برای یادگیری زمینه‌های دوطرفه با به حداکثر رساندن احتمال مورد انتظار بر روی همه جایگشت‌های ترتیب فاکتورسازی توالی ورودی، آموزش داده شده است. در این بخش پرسش q و c_{i^*} با استفاده از [SEP] به یک دیگر الحاق می‌شوند، و در اختیار XLNet قرار می‌گیرند که موقعیت‌های پاسخ را که با c_{i^*} شروع و ختم می‌شود، پیش‌بینی می‌کند. کلمات بین موقعیت شروع و پایان پیش بینی شده را به عنوان پاسخ a_k استخراج می‌شوند. در این مدل از یک XLNet از پیش آموزش‌دیده استفاده می‌شود که بر روی جفت‌های پرسش و پاسخ مبتنی بر دانش مجموعه‌داده AQUA تنظیم می‌شوند [12].

۳. مجموعه‌داده

در این کار از مجموعه‌داده AQUA برای پاسخ به پرسش دیداری در تصاویر هنری استفاده می‌کنیم. مجموعه‌داده‌ی AQUA برگرفته از مجموعه‌داده SemArt [11] می‌باشد. مجموعه‌داده SemArt شامل نقاشی‌ها و نظرات مرتبط است که نظرات در آن بلوک‌هایی از متن هستند. این نظرات به عنوان دانش عمل می‌کنند. برای نشان دادن پتانسیل‌های فناوری‌های هوش مصنوعی برای درک نقاشی‌ها، جفت‌های پرسش و پاسخ موجود در این مجموعه‌داده دارای دو حالت دیداری و مبتنی بر دانش خارجی هستند و از روش‌هایی متناسب با این حالات برای تولید پرسش‌ها استفاده شده است. در این مجموعه‌داده از دو روش برای تولید پرسش‌های دیداری استفاده شده است. اولین مورد، iQAN است که بر روی نسخه‌ی دوم مجموعه‌داده VQA [20] آموزش دیده است، که یک تصویر و یک کلمه پاسخ را به عنوان ورودی می‌گیرد و با استفاده از یک مدل شبکه عصبی یک پرسش ایجاد می‌کند. در مورد دوم، از Pythia [21] برای ایجاد یک عنوان برای هر نقاشی استفاده می‌شود و با استفاده از تکنیک TQG مبتنی بر قانون [۲۲]، هر عنوان تولید شده را به یک جفت پرسش و پاسخ تبدیل می‌کند.

برای ایجاد پرسش‌هایی که برای پاسخ به آن‌ها نیازمند دانش در مورد هنر هستیم، در مجموعه‌داده AQUA از روش‌های TQG[6] استفاده شده است. در این مجموعه‌داده چندین رویکرد TQG مورد امتحان قرار گرفته است، به عنوان مثال، رویکردهای مبتنی بر قانون و رویکردهای عصبی [23]. جدول ۱ آمار مجموعه‌داده AQUA را نشان می‌دهد.

جدول ۱: **جزئیات آماری مجموعه‌داده AQUA**

	آزمایش	اعتبارسنجی	آموزش	تعداد
زوج‌های QA	۴۹۱۲	۵۱۲۴	۶۹۸۱۲	
دیداری	۱۲۷۰	۱۵۰۷	۲۹۵۶۸	
دانشی	۳۶۴۲	۳۶۱۷	۴۰۲۴۴	
طول پرسش	۹/۴۱	۹/۲۱	۸/۸۲	
دیداری	۶/۵۱	۶/۵۰	۶/۵۳	
دانشی	۱۰/۴۳	۱۰/۳۳	۱۰/۵۰	
طول پاسخ‌ها	۳/۸۵	۳/۶۸	۳/۱۳	
دیداری	۱/۰۰	۱/۰۰	۱/۰۰	
دانشی	۴/۸۵	۴/۷۹	۴/۶۹	

۴. نتایج

در این بخش نتایج آزمایش‌های به دست آمده در شاخه‌های مختلف را بررسی می‌کنیم. عملکرد کار ما با تطابق دقیق[7] (EM) اندازه‌گیری می‌شود، یعنی درصد پیش‌بینی‌هایی که دقیقاً با داده مرجع مطابقت دارند.

۴-۱- *نتایج انتخاب ماهیت*

همانند مدل پایه [12]، در بخش انتخاب ماهیت توانستیم با دقت ۹۹/۶ درصد پرسش‌های دیداری را از پرسش‌های مبتنی بر دانش تمییز دهیم. از آن‌جایی که پرسش‌های دیداری و مبتنی بر دانش در مجموعه‌داده از روش‌های مختلفی ایجاد شده‌اند، تشخیص آن‌ها برای طبقه‌بندی کننده نسبتاً آسان است. در جدول ۲ ماتریس خطای انتخاب کننده ماهیت پرسش‌ها را ملاحظه می‌کنیم.

جدول ۲: **ماتریس خطای انتخاب کننده ماهیت**

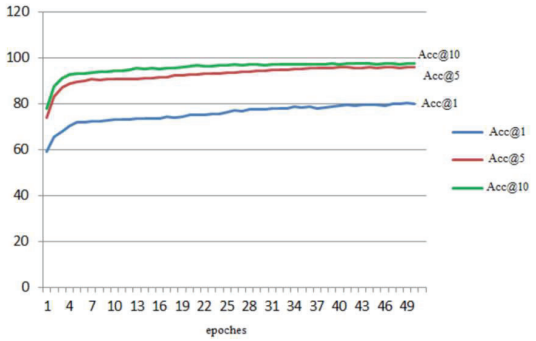
برچسب	پیش‌بینی	
	دیداری	دانشی
دیداری	۱۲۶۹	۱
دانشی	۱۷	۳۶۲۵

۴-۲- *نتایج شاخه دیداری*

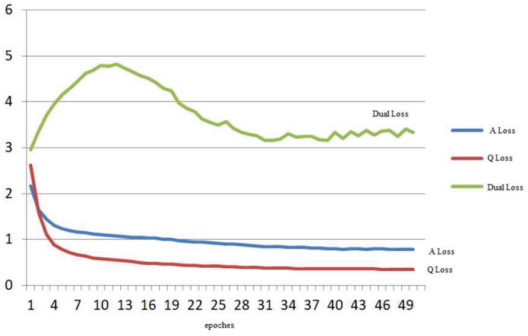
در تقسیم آموزشی شاخه‌ی پاسخ به پرسش دیداری، توانستیم به ۱۰۱۵ پرسش از ۱۲۸۶ پرسش پاسخ صحیح بدهیم و به دقت ۷۸/۹۲ درصد در شاخه‌ی دیداری و وظیفه‌ی پاسخ به پرسش دیداری برسیم. شکل ۱ نمودار دقت بر اساس مبناهای Acc@k بر

[5] Autoregressive
[6] Textual Question Generation
[7] Exact Match

روی تقسیم اعتبار سنجی را نشان می‌دهد، برای مثال Acc@5 به معنی این است که پاسخ صحیح در بین پنج پاسخ انتخاب شده می‌باشد. شکل ۲ نیز نمودار میزان اتلاف بر روی تقسیم اعتبارسنجی را نمایش می‌دهد. لازم به ذکر است که این دقت بهترین دقت به دست آمده توسط مدل پیشنهادی ما می‌باشد که با تابع فعال‌سازی tanh [24] و تعداد نگاه اجمالی ۲ به دست آمده که نشان دهنده بهینه بودن ۲ نگاه اجمالی به تصویر در حین آموزش برای مکانیسم توجه و تابع فعال‌سازی tanh به همراه آن می‌باشد.



شکل ۱: نمودار دقت بر روی تقسیم اعتبارسنجی.



شکل ۲: نمودار میزان اتلاف بر روی تقسیم اعتبار سنجی.

دقت به دست آمده در حالتی که تابع فعال‌سازی ReLU [25] و تعداد نگاه اجمالی برابر با ۲ باشد برابر با ۷۸/۱۴ درصد، در حالتی که تابع فعال‌سازی ReLU و تعداد نگاه اجمالی ۴ باشد برابر با ۷۸/۴۶ درصد و در حالتی که تابع فعال‌سازی tanh و نگاه اجمالی ۴ باشد برابر با ۷۷/۹۹ درصد می‌باشد. با توجه به نتایج به دست آمده، تابع فعال سازی tanh نسبت به ReLU میزان اتلاف بهینه‌تری دارد و علارغم نتیجه‌ی ضعیف‌تر با ۴ نگاه اجمالی، تابع tanh با ۲ نگاه اجمالی بهینه‌ترین نتیجه را در شاخه پاسخ به پرسش‌های دیداری در مدل پیشنهادی، ارائه می‌دهد.

۳-۴- *نتایج شاخه‌ی مبتنی بر دانش خارجی*

برای بازیابی دانش خارجی عملکرد به صورت حضور در k اندازه‌گیری می‌شود (R@k)، یعنی درصد جفت‌های پرسش و پاسخ که نظر اصلی مربوط به آن‌ها در k موقعیت‌های برتر رتبه‌بندی شده است. بازیابی دانش خارجی دو مرحله‌ای مورد استفاده در مدل پایه به بالاترین عملکرد دست می‌یابد. به طور خاص، نوع کامل (یعنی TF-IDF + PP + n-grams، که در آن PP مخفف پیش‌پردازش است) در مرحله اول نظرات اصلی را در ۱۰ موقعیت برتر برای بیش از ۹۰ درصد جفت‌های پرسش و پاسخ رتبه بندی می‌کند و در مرحله دوم با استفاده از یک مدل مبتنی بر BERT به رتبه بندی مجدد نظرات می‌پردازد. در کل در شاخه‌ی برمبنای دانش در تقسیم آزمایشی پرسش‌های برمبنای دانش توانستیم برای ۱۷۳۰ پرسش از ۳۶۲۶ پاسخ صحیح را بیابیم. این یعنی ۴۷/۷ درصد پاسخ دقیق. نتایج این شاخه در جدول ۳ آورده شده است.

جدول ۳: **نتایج شاخه‌ی مبتنی بر دانش خارجی**

مرحله بازیابی نظرات	حضور در اولین نظر برتر	۷۷/۱۶ درصد
	حضور در ۵ نظر برتر	۸۸/۰۸ درصد
	حضور در ۱۰ نظر برتر	۹۰/۹۵ درصد
مرحله پیش‌بینی پاسخ	تعداد پاسخ‌های صحیح	۱۷۳۰
	تعداد پرسش‌های دانشی	۳۶۲۶
	دقت نهایی	۴۷/۷۱ درصد
	نمره f1	۵۸/۵۲۲

۴-۴- *نتایج نهایی*

به طور کلی در تقسیم آزمایشی AQUA، ۴۹۱۲ پرسش داریم. در شاخه‌ی دیداری توانستیم با جایگزینی ماژول توجه MUTAN با ماژول توجه MLB به ۱۰۱۵ پرسش دیداری پاسخ دهیم و دقت را نسبت به شاخه دیداری مدل پایه از ۷۷/۷۶ به ۷۸/۹۲ درصد بهبود دهیم. در شاخه مبتنی بر دانش نیز به ۱۷۳۰ پرسش پاسخ دادیم. در مجموع دو شاخه با استفاده از مدل پیشنهادی به دقت ۵۵/۸۸ رسیدیم. جدول ۴ نتایج نهایی به دست آمده توسط مدل پیشنهادی را نشان می‌دهد.

جدول ۴: **نتایج نهایی به دست آمده توسط مدل پیشنهادی**

دقت EM	پرسش‌ها	پرسش‌های شاخه	پاسخ صحیح	شاخه
۰/۲۰۶۶	۴۹۱۲	۱۲۷۰	۱۰۱۵	دیداری
۰/۳۵۲۱	۴۹۱۲	۳۶۲۶	۱۷۳۰	دانشی
۰/۵۵۸۸	۴۹۱۲	۴۹۱۲	۲۷۴۵	نتیجه کل

۵-۴- مقایسه روش پیشنهادی با سایر روش‌ها

در جدول ۵ عملکرد مدل پیشنهادی را با برخی از مدل‌های دیگر بر روی تقسیم آزمایشی AQUA مقایسه می‌کنیم. همانطور که مشخص است، مدل پیشنهادی ما (Ours) با توجه به بهودهایی که در شاخه دیداری اعمال شدند، در مقایسه با مدل پایه (VIKING) و سایر مدل‌های مطرح عملکرد بهتری دارد. در این جدول P به معنی استفاده از نقاشی‌ها، K برای استفاده از دانش، Q برای پرسش‌ها و w/o به معنی عدم استفاده است. روش‌های LSTM [26]، BERT و XLNet تنها از پرسش‌ها برای پاسخ به پرسش‌های مجموعه‌داده استفاده می‌کنند. مدل BAN [27] از نقاشی‌ها و پرسش‌ها برای پاسخ‌دهی به سوالات استفاده می‌کند.

جدول ۵: **مقایسه دقت به دست آمده در مدل‌های مختلف**

روش	Q	P	K	EM
LSTM	✓	–	–	۰/۱۹۸
BERT	✓	–	–	۰/۱۹۴
XLNet	✓	–	–	۰/۱۹۳
BAN	✓	✓	–	۰/۲۲۴
VIKING w/o K	✓	✓	–	۰/۲۰۴
VIKING w/o P	✓	–	✓	۰/۳۵۲
VIKING full	✓	✓	✓	۰/۵۵۵
Ours w/o K	✓	✓	_	۰/۲۰۶
Ours w/o P	✓	_	✓	۰/۳۵۲۱
Ours full	✓	✓	✓	۰/۵۵۸۸

۵. نتیجه‌گیری

در این مقاله یک مدل پاسخ به پرسش دیداری به تصاویر هنری با استفاده از دانش خارجی ارائه شد. استفاده از مجموعه‌داده‌های با موضوع تخصصی از این جهت که به کاربردی‌تر شدن سیستم‌های پاسخ به پرسش دیداری در حوزه‌های مختلف کمک می‌کند، اهمیت دارد. برای این کار ما بهبودی بر روی مدل VIKING ارائه دادیم و نشان دادیم که مدل پیشنهادی ما در مقایسه با سایر مدل‌های مطرح و مدل پایه دقت بهتری را کسب می‌کند. توانستیم دقت را در شاخه‌ی دیداری از ۷۷/۷۶ به ۷۸/۹۲ درصد افزایش دهیم. در مجموع دو شاخه نیز توانستیم دقت را از ۵۵/۵ درصد به ۵۵/۸۸ درصد افزایش دهیم. با توجه به بیشتر بودن تعداد پرسش‌های مبتنی بر دانش در مجموعه‌داده AQUA در کارهای آینده می‌توانیم بر روی شاخه‌ی مبتنی بر دانش نیز تغییراتی اعمال کنیم و با استفاده از مدل‌های پیش آموزش بهینه‌تر دقت را در این شاخه نیز افزایش دهیم. استفاده از مدل‌های جایگزین iQAN در شاخه‌ی دیداری و بزرگ‌تر کردن مجموعه‌داده و افزایش تنوع پرسش‌ها در مجموعه‌داده نیز به بهبود نتایج در این حوزه کمک می‌کند.

مراجع

[1] L. Shamir, T. Macura, N. Orlov, D. M. Eckley, and I. G. Goldberg, "Impressionism, expressionism, surrealism," *ACM Transactions on Applied Perception*, vol. 7, no. 2, pp. 1–17, Feb. 2010.
[2] C. Johnson *et al.*, "Image processing for artist identification," *IEEE Signal Processing Magazine*, vol. 25, no. 4, pp. 37–48, Jul. 2008.
[3] D. Ma *et al.*, "From Part to whole: Who Is behind the painting?" 2017.
[4] G. Carneiro, Silva, A. D. Bue, and J. P. Costeira, "Artistic image classification: An analysis on the PRINTART database," 2012.
[5] W. R. Tan, C. S. Chan, H. E. Aguirre, and K. Tanaka, "Ceci n'est pas une pipe: A deep convolutional network for fine-art paintings classification," 2016.
[6] N. García, B. Renoust, and Y. Nakashima, "Context-aware embeddings for automatic art analysis," 2019.
[7] N. Huckle, N. García, and Y. Nakashima, "Demographic influences on contemporary art with unsupervised style embeddings," *ArXiv*, vol. abs/2009.14545, 2020.
[8] G. Carneiro, Silva, A. Del Bue, and J. P. Costeira, "Artistic image classification: An analysis on the PRINTART database," 2012.
[9] E. J. Crowley and A. Zisserman, "The state of the art: Object retrieval in paintings using discriminative regions," 2014.
[10] E. J. Crowley, O. M. Parkhi, and A. Zisserman, "Face Painting: querying art with photos," 2015.
[11] N. García and G. Vogiatzis, "How to read paintings: Semantic art understanding with multi-modal retrieval," 2018.
[12] N. Garcia *et al.*, "A dataset and baselines for visual question answering on art," *CoRR*, vol. abs/2008.12520, 2020.
[13] Y. Li, N. Duan, B. Zhou, X. Chu, W. Ouyang, and X. Wang, "Visual question generation as dual task of visual question answering," *CoRR*, vol. abs/1709.07192, 2017.
[14] Z. Yang, Z. Dai, Y. Yang, J. G. Carbonell, Ruslan Salakhutdinov, and Q. V. Le, "XLNet: Generalized autoregressive pretraining for language understanding," *CoRR*, vol. abs/1906.08237, 2019.
[15] J.-H. Kim, Kyoung Woon On, W. Lim, J. Kim, J. Ha, and B.-T. Zhang, "Hadamard product for low-rank bilinear pooling," *CoRR*, vol. abs/1610.04325, 2016.
[16] H. Ben-Younes, R. Cadène, M. Cord, and N. Thome, "MUTAN: Multimodal tucker fusion for visual question answering," *CoRR*, vol. abs/1705.06676, 2017.
[17] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," *CoRR*, vol. abs/1810.04805, 2018.
[18] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *CoRR*, vol. abs/1512.03385, 2015.
[19] Z. Dai, Z. Yang, Y. Yang, J. G. Carbonell, Q. V. Le, and Ruslan Salakhutdinov, "Transformer-xl: Attentive language models beyond a fixed-length context," *CoRR*, vol. abs/1901.02860, 2019.
[20] Y. Goyal, Tejas Khot, D. Summers-Stay, D. Batra, and D. Parikh, "Making the V in VQA matter: Elevating the role of image understanding in visual question answering," *CoRR*, vol. abs/1612.00837, 2016.
[21] A. Singh *et al.*, "MMF: A multimodal framework for vision and language research," 2020.
[22] M. Heilman and N. A. Smith, "Good question! Statistical ranking for question generation," pp. 609–617, Jun. 2010.
[23] X. Du, J. Shao, and C. Cardie, "Learning to ask: Neural question generation for reading comprehension," *CoRR*, vol. abs/1705.00106, 2017.
[24] W Malfliet, "The tanh method: a tool for solving certain classes of nonlinear evolution and wave equations," *Journal of Computational and Applied Mathematics*, vol. 164–165, pp. 529–541, 2004.
[25] Abien Fred Agarap, "Deep learning using rectified linear units (ReLU)," *CoRR*, vol. abs/1803.08375, 2018.
[26] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, pp. 1735–80, Dec. 1997.
[27] J.-H. Kim, J. Jun, and B.-T. Zhang, "Bilinear attention networks," *CoRR*, vol. abs/1805.07932, 2018.