مقاله پژوهشی

مجله ماشین‌بینایی و پردازش‌تصویر
سال دهم، شماره اول، بهار ۱۴۰۲

توصیف محتوای تصاویر به‌صورت خودکار با بکارگیری شبکه عصبی کپسولی و روش تعبیه‌سازی ELMo

شیما جوانمردی، علی‌محمد لطیف، محمدتقی صادقی

چکیده

توصیف محتوای تصویر به‌صورت خودکار توسط ماشین یک مشکل چالش‌برانگیز در بینایی کامپیوتر است و هدف آن تولید توضیحات قابل درک توسط کامپیوتر برای تصاویر می‌باشد. استفاده از شبکه‌های عصبی پیچشی (CNN) نقش مهمی در زمینه توصیف تصاویر ایفا کرده است. با این حال، در طول فرآیند تولید توصیف‌های مرتبط با تصویر دو چالش عمده برای CNN وجود دارد که عبارتند از: عدم توجه آنها به روابط و ساختارهای سلسله مراتبی مکانی بین اشیای درون تصویر، و عدم ثبات در مقابل تغییرات چرخشی تصاویر. به‌منظور رفع این چالش‌ها این مقاله با بهره‌گیری از یک شبکه کپسولی بهبود یافته، به توصیف محتوای تصویر با استفاده از پردازش زبان‌طبیعی می‌پردازد. شبکه کپسولی با درنظر گرفتن موقعیت مکانی اشیای درون تصویر نسبت به یکدیگر، اطلاعات مربوط به محتوای تصویر را ارائه می‌دهد. یک کپسول شامل مجموعه‌ای از نورون‌ها است که پارامترهای وضعیت اشیا درون تصویر مانند سایز، جهت، مقیاس و روابط اشیا نسبت به هم را در نظر می‌گیرند. این کپسول‌ها تمرکز ویژه‌ای بر استخراج ویژگی‌های معنادار برای استفاده در فرآیند تولید توضیحات مرتبط برای مجموعه‌ای معین از تصاویر دارند. آزمایش‌های کیفی روی مجموعه داده‌های MS-COCO با بهره‌گیری از شبکه کپسولی و روش تعبیه‌سازی ELMo، منجر به بهبود ۲ تا ۵ درصدی معیارهای ارزیابی شده، در مقایسه با مدل‌های زیرنویس تصویر موجود شده است.

کلید واژه ها

توصیف تصاویر، یادگیری عمیق، شبکه کپسولی، شبکه عصبی پیچشی، پردازش زبان طبیعی.

۱- مقدمه

توصیف تصویر به‌صورت خودکار یک چالش در بینایی کامپیوتر با کاربردهای بسیاری در زمینه‌های مختلف تحقیقاتی است. هدف این موضوع، تولید توصیف‌های غنی برای تصاویر و ارائه توضیحات قابل درک برای انسان با توجه به محتوای تصاویر می‌باشد [1].

در این مقاله، هدف اصلی، بررسی مدل‌ها و ارائه روش‌هایی برای درک بهتر دامنه بصری تصاویر توسط سیستم‌های کامپیوتری می‌باشد به‌گونه‌ای که با بهره‌گیری از پردازش زبان طبیعی[1] به توصیف محتوای تصویر و ایجاد ارتباط بین دو حوزه پردازش تصویر و متن می‌پردازد.

با افزایش حجم تصاویر دیجیتال، منابع تصویری مختلف در اینترنت مانند مقالات خبری، تبلیغات، وبلاگ‌ها و موارد مشابه در دسترس می‌باشد. به‌طورکلی بیننده یک تصویر بایستی محتوا آن را خودش تفسیر کند و اکثر تصاویر فاقد توضیح هستند. در نظر گرفتن معادل ظاهری بین یک تصویر و هزاران کلمه چالش برانگیز است. به‌علاوه، این برای فردی با ظرفیت ادراکی معمولی ساده

[۱] دانشجوی دکتری هوش مصنوعی، دانشگاه یزد.
رایانامه: sh.javanmardii@gmail.com
[۲] دانشکده مهندسی کامپیوتر، دانشگاه یزد
رایانامه: alatif@yazd.ac.ir
[۳] دانشکده مهندسی برق، دانشگاه یزد.
رایانامه: m.sadeghi@yazd.ac.ir

نویسنده مسئول: علی‌محمد لطیف

dorl.net/dor/20.1001.1.23831197.1402.10.1.6.4

[1] Natural Language processing (NLP)

نیست. تفسیر محتوای حجم زیادی از تصاویر به‌صورت دستی کاری چالش برانگیز و زمان‌بر است. از این رو، در این مقاله یک روش تولید توصیف محتوای تصویر به‌صورت خودکار مبتنی بر شبکه‌های عصبی پیچشی[1] و شبکه‌های عصبی بازگشتی[2] ارائه خواهد شد که به تفسیر محتوای تصویر می‌پردازد.

توصیف محتوای تصاویر برای بسیاری از زمینه‌های کاربردی مانند حاشیه‌نویسی خودکار تصویر [4]–[2]، درک صحنه و بازیابی تصویر از موتورهای جستجو ضروری است. توصیف تصویر به‌طور موثر به افراد نابینا اجازه می‌دهد تا محیط اطراف خود را درک کنند. این حوزه تحقیقاتی دارای موارد استفاده متعددی از جمله زیست پزشکی، تجارت و آموزش است و به‌طور عمده در کتابخانه‌های دیجیتال و موتورهای جستجوی وب مورد استفاده قرار می‌گیرد[5] .

شکل ۱ نمونه‌ای از نتایج ارزیابی شده توسط مدل پیشنهادی در این مقاله را نشان می‌دهد. در این مدل، یک سیستم رمزگذار-رمزگشا محتوای تصاویر را به زبان طبیعی توصیف می‌کند. استخراج ویژگی‌های معنایی از اشیای درون تصاویر و همچنین روابط معنادار بین آن ویژگی‌ها در بخش رمزگذاری شبکه انجام می‌شود [6]. خروجی رمزگذار دنباله‌ای از اعداد صحیح است که حاوی اطلاعات مربوط به محتوای بصری و مفاهیم سطح بالا در هر تصویر می‌باشد. این بردار که طول ثابتی دارد به‌عنوان ورودی به رمزگشا داده می‌شود تا پردازش زبان طبیعی را برای توصیف تصویر انجام دهد.

توصیف پیش‌بینی شده:

A man wearing a suit and tie and red hat with a silver buckle.

توصیف اصلی:

A man wearing a hat and a tie.

شکل (۱) : نمایی کلی از یک شبکه تولید توصیف برای تصویر

عملکرد مدل‌های تولید توصیف برای تصویر ارتباط نزدیکی با کیفیت ویژگی‌های استخراج‌شده از تصاویر و قدرت مدل زبانی در تولید توضیحات دقیق و معنادار مرتبط با محتوای تصویر دارد. در نظرگرفتن روابط معنایی بین اشیای شناسایی شده در تصویر امری ضروری در فرآیند توصیف تصویر به‌شمار می‌رود. بااین‌حال، در مدل‌های اخیر تولید توصیف برای تصویر، شناسایی اشیای درون تصویر (به عنوان فاعل جمله توصیفی) و نحوه یافتن تعامل و ارتباط میان اشیا (به عنوان فعل جمله توصیفی) نیز یک چالش بزرگ می‌باشد. از طرف دیگر به محتوای بصری تصاویر به‌تنهایی نمی‌توان اتکا کرد و ویژگی‌های مفهومی تصاویر نیز نقش مهمی در این زمینه ایفا می‌کند. در سال‌های اخیر، بسیاری از روش‌های مختلف توصیف تصویر از الگوریتم‌های یادگیری عمیق برای کنترل پیچیدگی و رسیدگی به این چالش‌های فرآیند شرح تصویر استفاده می‌کنند. این رویکردها هنوز به‌طورکامل قادر به تولید توصیف‌های واقع‌گرایانه که تمام مفاهیم تصویر را در بر می‌گیرد، نمی‌باشند.

بسیاری از مدل‌های ارائه‌شده برای توصیف محتوای تصویر، به‌طور عمده از شبکه‌های عصبی پیچشی (CNN) به‌عنوان استخراج‌کننده ویژگی تصویر استفاده می‌کنند. CNN نمی‌تواند به‌صورت قابل‌توجهی اشیای برجسته تصویر و روابط آنها را برای ایجاد یک توصیف معنادار برای تصویر شناسایی کند. در سال‌های اخیر، بسیاری از روش‌های ایجاد توصیف برای تصاویر از الگوریتم‌های یادگیری عمیق برای کنترل پیچیدگی و چالش‌های فرآیند توصیف محتوای تصویر استفاده کرده‌اند [12]–[7].

بنابراین، انگیزه ما در این مقاله توسعه روشی جدید است که می‌تواند به‌گونه موثری روابط میان اشیای درون تصویر را در طول فرآیند توصیف تصویر درنظر بگیرد و توصیف‌های معنادارتری باتوجه به محتوای تصویر تولید کند. از این رو، در این مقاله، یک مکانیزم رمزگذار-رمزگشا برای توصیف محتوای تصاویر پیشنهاد داده می‌شود که با توسعه یک ساختار جدید از یک شبکه کپسولی (CapsNet)به مقابله با این چالش‌ها می‌پردازد [13]. مدل پیشنهادی، با در نظر گرفتن روابط بین اشیای درون تصویر، توصیف‌های معناداتر و متنوع‌تری را برای تصویر از طریق یک مدل زبان ایجاد می‌کند.

CapsNet می‌تواند با درنظرگرفتن بافت تصویر، به‌طور موثر کمبودهای CNN را با شناسایی سطوح همپوشانی شده شناسایی کند [14]. CapsNet قادر است نواحی برجسته‌تر تصویر که حاوی اشیای درون تصویر هستند را تشخیص داده و ویژگی‌های هندسی اشیای درون تصویر، مانند جهت، اندازه، مقیاس اشیا نسبت به هم را برای هر تصویر ورودی با برداری از اعداد معنادار نشان دهد. این جنبه از CapsNet با CNN در تضاد است زیرا فقدان ویژگی‌های تغییرناپذیری محلی باعث ایجاد تغییرات بیش از حد در خروجی‌های ایجاد شده توسط CNN می‌شود [15]. علاوه بر این، مدل پیشنهادی این مقاله در قسمت رمزگشایی شبکه با بکارگیری روش تعبیه‌سازی ELMo، به مدل‌سازی ویژگی‌های پیچیده زبانی مانند نحو و معنای جملات و چگونگی تغییر این کاربردها در متون زبانی متنوع، می‌پردازد. هدف این ایده، بهبود توانایی مدل زبانی در تولید متن‌های توصیفی با دقت و کیفیت بالاتری می‌باشد.

عملکرد مناسب مدل پیشنهادی بر روی مجموعه داده MS-COCO با مقیاس بزرگ نشان داده شده است. این درحالیست که

[1] Convolutional Neural Network
[2] Recurrent Neural Network

در پژوهش اخیر ارائه‌شده در این حوزه عنوان شده است که پیاده‌سازی شبکه کپسولی بر روی مجموعه داده MS-COCO به دلیل بزرگی مقیاس و پیچیدگی داده‌ها نیازمند حجم بالایی از محاسبات و منابع پردازشی می‌باشد [16]. لذا در این پژوهش شبکه کپسولی بر روی مجموعه دادگان MS-COCO پیاده‌سازی شده است.

در این مقاله، در بخش دوم مروری بر تحقیقات گذشته در حوزه توصیف تصویر مورد بررسی قرار می‌گیرد. سپس در بخش سوم، مروری بر فرآیند یادگیری عمیق انجام خواهد شد. در ادامه شبکه‌های عصبی پیچشی، شبکه کپسولی[1] وشبکه بازگشتی معرفی می‌شود. در بخش چهارم، فرآیند توصیف تصاویر که موضوع اصلی مقاله را به‌خود اختصاص می دهد تحلیل و بررسی می‌شود. بخش پنجم، بیان مسئله و معرفی مدل‌های پیاده‌سازی شده در مقاله معرفی می‌شود. در بخش ششم نتایج ارزیابی شده بررسی و تحلیل خواهند شد. در نهایت، در بخش هفتم نتیجه‌گیری کلی از رویکرد پیشنهادی در مقاله ارائه می‌شود.

۲- مروری بر کارهای گذشته

فرآیند توصیف تصاویر دارای سه رویکرد رایج می‌باشد. در حالت اول از یکسری قالب برای توصیف تصاویر استفاده می‌شود که این قالب‌ها برمبنای اشیای درون تصویر و روابط میان آنها بیان شده است. حالت دوم بر اساس بازیابی توصیف‌های مشابه از پایگاه‌های داده و اصلاح آنها مطابق با محتوای تصویر ایجاد می‌شوند. هردو این رویکردها دارای محدودیت در توصیف تصاویر متنوع می‌باشند و از انعطاف کافی برخوردار نبوده و بسیار وابسته به داده‌های یادگیری شده توسط شبکه می‌باشند. رویکرد سوم، توصیف تصاویر بر مبنای مدل‌های یادگیری عمیق ارائه شده است که دارای قابلیت بسیار بالایی در توصیف محتوای تصویر می‌باشد.

رویکردهای ارائه‌شده برای فرآیند توصیف تصاویر شامل یک فاز تعمیم برای حذف اطلاعات غیرمهم مربوط به تصویر، می‌باشند. بسیاری از مدل‌های کنونی از شبکه‌های عصبی استفاده می‌کنند که این الهام‌گرفته از مزایای شبکه‌های عصبی بازگشتی در فرآیند ترجمه متون میان زبان‌های مختلف می‌باشد که دنباله‌ای از کلمات را به‌عنوان ورودی دریافت می‌کند و طی فرآیند توصیف تصویر، ورودی را به یک بردار با طول ثابت تبدیل می‌کند و در نهایت بردار رمزشده را به‌صورت یک توالی دیگر از کلمات در خروجی رمزگشایی می‌کند مشابه این عملیات در [17] ارائه شده است.

از معماری رمزگذار-رمزگشا نیز در فرآیند توصیف تصاویر می‌توان بهره برد. زیرا این مسائل را می‌توان به‌صورت ترجمه تصویر به متن تفسیر کرد. در این مسائل بخش رمزگذار، مدلی شبیه به شبکه‌های عصبی پیچشی است که عملکرد مناسبی در فرآیند طبقه‌بندی تصاویر ارائه داده‌اند [18]. بخش رمزگشا نیز مشابه با مدل ترجمه ماشین است که در آن از شبکه‌های عصبی بازگشتی مانند LSTM[2] استفاده می‌شود. بسیاری از پژوهش‌های اخیر، در فرآیند توصیف تصاویر برمبنای یکسری مدل‌های آماری ایجاد شده‌اند که در آنها با توجه به ویژگی‌های استخراج شده از نواحی مختلف تصویر به نام‌گذاری و توصیف آن نواحی می‌پردازند [21]–[19].

بسیاری از رویکردهای ارائه شده به‌منظور توصیف محتوای تصاویر، علاوه‌بر اسامی اشیا، از حروف اضافه و صفت‌های مقایسه‌ای نیز برای بیان روابط بین اشیا بهره می‌برند. به‌طور مثال در [22] فرهادی و همکارانش با بهره‌گیری از یک مدل زبانی رایج، با استفاده از جملات به بیان محتوای تصاویر می‌پردازند. در پژوهش آنها مدلی ارائه می‌شود که تصاویر و جملات را به فضای معنایی مشترک نگاشت می‌کند. در این مدل برمبنای اطلاعات سه‌گانه (شی، عمل، صحنه) از هرتصویر و هر جمله استخراج می‌شود و سپس برمبنای این اطلاعات با استفاده از یک جمله به‌صورت پرس‌وجو به بازیابی تصاویر می‌پردازد و یا برعکس. در [23] Fei-Fei و همکارانش مدلی ارائه می‌دهند که در آن با اعمال روش آنالیز همبستگی استاندارد مبتنی بر کرنل[3] بر ویژگی‌های استخراج‌شده از تصاویر و متون، بخش‌های مختلف تصویر و کلمات را به یک فضای معنایی مشترک نگاشت می‌کنند. در [24] رویکردی ارائه شده است که در آن با استفاده از یک موتور تجزیه سلسله‌مراتبی به پردازش تصاویر می‌پردازد سپس بر مبنای محتوای تولیدشده از تصویر و با بهره‌گیری از یک فرآیند واسط به تولید متن به‌منظور توصیف تصاویر می‌پردازد. در رویکردی مشابه، در [25] از یک گراف AND-OR به‌منظور توصیف رویدادهای ورزشی در تصاویر ویدئویی استفاده شده است. در [26] مدلی ارائه می‌شود که در آن به شناسایی اشیا درون تصاویر می‌پردازد و سپس آنها را با جملات ایجاد شده از n-گرم‌های از پیش آموزش‌دیده موجود در جدول حاوی عبارات پرتکرار، ترکیب می‌کند. در [27] با بازیابی اشیا درون تصاویر و ترکیب آنها با جملات دست‌نویس ایجاد شده توسط افراد، تصاویر توصیف می‌شوند. در [28] رویکردی ارائه می‌شود که در آن با استفاده از فرآیندهایی چون تشخیص اشیا، صحنه و یا ویژگی‌های فضایی تصاویر، اطلاعات هر تصویر استخراج شده و به‌منظور توصیف محتوای تصویر به قالب‌های ثابت جملات تزریق می‌شود. در نهایت در [29] و به‌طور مشابه در [30] کلمات احتمالی توصیف‌کننده تصاویر بر مبنای تشخیص اشیا درون تصویر و ایجاد توصیف با بکارگیری درختی منطبق با قواعد نحوی، تخمین زده می‌شوند.

در بخش بعد به معرفی روش‌های یادگیری عمیق استفاده شده در این مقاله می‌پردازیم.

[1] Capsule Network

[2] Long-Short Term Memory

[3] Kernelized Canonical Correlation Analysis

۳- یادگیری عمیق

طی سالهای اخیر، یادگیری عمیق بصورت گسترده در حوزه بینایی کامپیوتر مورد مطالعه قرار گرفته است. یادگیری عمیق روشی مبتنی بر شبکه‌های عصبی است و به‌عنوان زیرمجموعه‌ای از یادگیری ماشین به‌شمار می‌رود. معماری‌های یادگیری عمیق متنوعی وجود دارد که از جمله‌ی آنها می‌توان به شبکه عصبی پیچشی [۱۶]، شبکه باور عمیق و شبکه عصبی بازگشتی [۱۸] اشاره کرد. در ادامه جزئیات این شبکه‌ها بررسی خواهد شد.

۳-۱- شبکه‌های عصبی پیچشی

شبکه‌های پیچشی یکی از مهمترین مدل‌های یادگیری عمیق هستند که آموزش لایه‌های تشکیل دهنده آنها با استفاده از یک روش قدرتمند صورت می‌گیرد [31]. به‌طور کلی یک شبکهCNN از سه لایه اصلی پیچشی[1]، ادغام[2] و تمام‌متصل[3] تشکیل می‌شود. در هر شبکه عصبی پیچشی دو مرحله، (۱) انتشار به جلو[4] و (۲) پس انتشار[5] برای آموزش وجود دارد. در مرحله اول تصویر ورودی به شبکه تغذیه می‌شود و این عمل ضرب نقطه‌ای بین ورودی و پارامترهای هر نورون و درنهایت اعمال عملیات پیچشی در هر لایه بوده و سپس عمل خروجی شبکه محاسبه می‌شود. در این مرحله به‌منظور تنظیم پارامترهای شبکه و یا به عبارت دیگر همان آموزش شبکه، از نتیجه خروجی به‌منظور محاسبه میزان خطای شبکه استفاده می‌شود. برای اینکار خروجی شبکه را با استفاده از یک تابع خطا[6] با پاسخ صحیح مقایسه کرده و به این صورت میزان خطا محاسبه می‌شود. در مرحله بعد، بر اساس میزان خطای محاسبه شده، مرحله پس‌انتشار آغاز می‌شود. در این مرحله گرادیان هر پارامتر با توجه به قاعده زنجیره[7] محاسبه می‌شود و تمامی پارامترها با توجه به تاثیری که بر خطای ایجاد شده در شبکه دارند تغییر پیدا می‌کنند. بعد از بروزرسانی پارامترها، انتشار به جلو شروع می‌شود. بعد از تکرار تعداد مناسبی از این مراحل آموزش شبکه پایان می‌یابد.

در رویکرد ارائه‌شده در این مقاله در فاز رمزگذاری از دو شبکه پیچشی و کپسولی به‌منظور استخراج ویژگی معنادار از تصاویر استفاده می‌شود که در بخش بعد به معرفی شبکه کپسولی و بررسی ساختار آن می‌پردازیم.

۳-۲- شبکه کپسولی:

در این مقاله شبکه کپسولی را از ابتدا به عنوان استخراج کننده ویژگی‌های بصری آموزش می‌دهیم. شبکه کپسولی [13] برای غلبه بر اشکالات موجود در CNN معرفی شده است. برخلاف یک شبکه پیچشی، کپسول‌ها در این شبکه اطلاعات جامعی درباره مکان و وضعیت یک شی و موقعیت نسبت به دیگر اشیا در تصویر را ذخیره می‌کنند.

هینتون و همکارانش در [13] مدعی شدند که صرف‌نظر از قابلیت‌های بالای شبکه‌های عصبی پیچشی، این شبکه دو عیب عمده دارد: ۱- ثبات در برابر چرخش و ۲- استفاده از لایه ادغام. اولین مورد باعث نارسایی در تشخیص روابط فضایی بین اشیا می‌شود و دومین مورد به دلیل انتخاب سلول با حداکثر مقدار در ناحیه از پیکسل‌های تصویر، باعث از دست رفتن اطلاعات حاصل از تصویر می‌شود. بنابراین صبور و همکاران در [13] شبکه کپسولی را برای رسیدگی به مسائل فوق پیشنهاد کردند.

این شبکه دارای یک لایه پیچشی و دو لایه کپسولی است و خروجی را با استفاده از سه لایه کاملا متصل می‌سازد. در مدل پیشنهادی ما در این مقاله به‌منظور استخراج ویژگی‌های بامعناتر از تصاویر با توجه به پیچیدگی مجموعه داده ورودی، با افزایش عمق شبکه از سه لایه پیچشی برای شبکه کپسولی استفاده کردیم. لایه‌های ادغام در شبکه کپسولی با مکانیزمی به نام "مسیریابی براساس توافق" جایگزین می‌شوند. بر اساس این مکانیزم، خروجی هرکپسول در سطح پایین تنها در صورتی به کپسول‌های والد در سطح بالاتر ارسال می‌شود که ویژگی‌های آنها وابستگی داشته باشد. شکل ۲ ساختار یک کپسول و نحوه هدایت داده‌ها بین کپسول‌های سطح پایین و سطح بالاتر را نشان می دهد.

در شکل ۲ بخش الف، هر کپسول، والد مناسب را در لایه بعدی در طول روش مسیریابی پویا پیدا می‌کند تا خروجی خود را به آن کپسول‌های لایه بالا ارسال کند. ورودی و خروجی کپسول ها بردار است. با در نظر گرفتن u_i به عنوان بردار پیش‌بینی کپسول i و $u_{j|i}$ به عنوان خروجی کپسول والد j در سطح بالاتر با ضرب u_i در ماتریس وزنی w_{ij} محاسبه می‌شود:

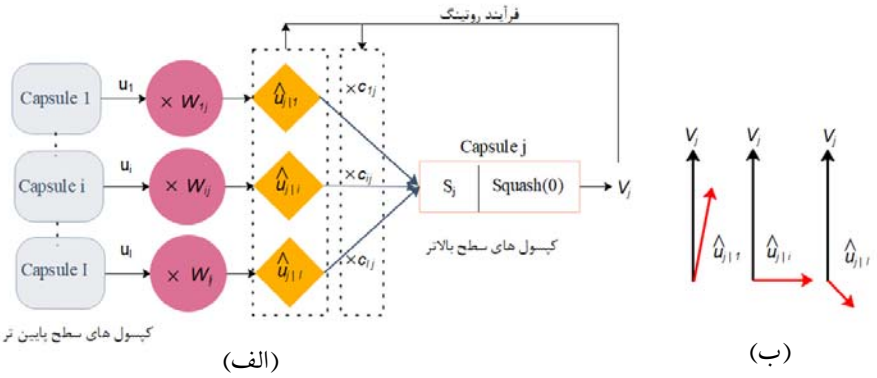
$$\hat{u}_{i|j} = w_{ij} . u_i \qquad (1)$$

طول u_i نشان‌دهنده احتمال پیش‌بینی یک شی در تصویر حتی پس از تغییر زاویه دید است. جهت u_i نشان‌دهنده چندین ویژگی آن شی، مانند اندازه و موقعیت آن است. مجموع وزنی $\widehat{u_{j|i}}$ و یک ضریب جفت میانی c_{ij}, به‌عنوان بردار ورودی کل به کپسول j با تابع ارائه شده در رابطه ۲ محاسبه می‌شود:

$$s_j = \sum_i c_{ij}\, \widehat{u_{j|i}} \qquad (2)$$

در این رابطه ضریب جفت c_{ij}، احتمال ویژه کلاس است که پس از مسطح کردن بردارها محاسبه می‌شود و توسط یک تابع سافت-مکس مسیریابی، به‌صورت رابطه ۳ محاسبه می‌شود:

[1] Convolution
[2] Pooling
[3] Fully Connected
[4] Feed Forward
[5] Backpropagation
[6] Loss Function
[7] Chain Rule



(الف) (ب)

شکل(۲) : ساختار یک شبکه کپسولی

$$c_{ij} = \frac{\exp(b_{ij})}{\sum_k \exp(b_{ik})} \quad (3)$$

در این رابطه b_{ij} لگاریتم احتمال اتصال بین کپسول i و j است. همانطور که در بخش (ب) شکل ۲ نشان داده شده است، مقدار c_{ij} زمانی افزایش می‌یابد که کپسول‌های سطح پایین‌تر و سطح بالاتر با پیش‌بینی‌های آنها سازگار باشند و زمانی کاهش می‌یابد که ناسازگار باشند. این پارامتر در روش مسیریابی، با مقدار صفر مقداردهی اولیه می‌شود. تابع غیرخطی اسکواش[1] با اعمال بر بردار ورودی در این شبکه جایگزین تابع فعال‌سازی ReLU شده و با رابطه ۴ به صورت زیر محاسبه می‌شود:

$$v_j = \frac{\|s_j\|^2}{1+\|s_j\|^2} \frac{s_j}{\|s_j\|} \quad (4)$$

در این رابطه s_j بردار ورودی و v_j خروجی نرمال شده بین صفر و یک است. لگاریتم احتمال همراه با مکانیزم مسیریابی با محاسبه توافق بین v_j به‌عنوان خروجی کپسول j در لایه بالا و $\hat{u}_{i|j}$ به‌عنوان بردار پیش‌بینی بروز می‌شود. تابع ضرر شبکه برای هر کپسول k به‌صورت رابطه ۵ محاسبه می‌شود:

$$L_k = T_k \max(0, m^+ - \|V_k\|)^2 + \lambda (1 - T_k) max(0, \|V_k\| - m^-)^2 \quad (5)$$

در این رابطه L_k نشان‌دهنده میزان خطا برای یک پیش‌بینی است، T_k پارامتری است که در صورت وجود کلاس k برابر با یک است.

[1] Squash

در غیر این صورت، برابر صفر خواهد بود. پارامترهای m^+ و m^- مشابه [13] برای کنترل محدوده احتمال روی ۰/۱ و ۰/۹ تنظیم می‌شوند. پارامتر λ برای کنترل وزن پایین وزن‌های اولیه برای کلاس‌های غایب استفاده می‌شود. $||\ .\ ||$ در تمام معادلات نشان دهنده نرم $L2$ است.

در نسخه بهبودیافته شبکه کپسولی، اندازه تصویر ورودی به شبکه کپسولی ۳×۲۲۹×۲۲۹ درنظر گرفته‌شده‌است. معماری متفاوت شبکه کپسولی، آن را در مقایسه با CNN متمایز می‌کند. به‌جز لایه‌های ورودی و خروجی، شبکه کپسولی از لایه کپسول اولیه و لایه کپسول طبقه‌بندی تشکیل شده است. خروجی کپسول طبقه‌بندی به رمزگشا ارسال می‌شود. شبکه با بازسازی تصویر ورودی از کپسول‌های خروجی با به حداقل رساندن تلفات بازسازی به‌عنوان یک روش منظم‌سازی در رمزگشا، از برازش بیش از حد جلوگیری می‌کنند [32]. فرضیه ما در این مقاله عملکرد مناسب این شبکه در استخراج ویژگی‌های موثر از تصاویر با در نظر گرفتن روابط میان اشیای درون تصویر می‌باشد. اطلاعات استخراج شده از بخش رمزگذار به شبکه بازگشتی برای تولید توصیف مرتبط با محتوای تصویر ارسال می‌شود. در بخش بعد ساختار این شبکه بررسی می‌شود.

۳-۳- شبکه‌های عصبی بازگشتی

بسیاری از شبکه‌های عصبی همچون شبکه‌های پیشرو دارای مدل‌های قدرتمندی هستند اما تنها برای مسائلی کاربرد دارند که ورودی‌ها و خروجی‌ها دارای ابعاد ثابتی هستند. این یک نقص جدی محسوب می‌شود، زیرا بسیاری از مسائل دنیای واقعی به‌صورت دنباله‌ای با طول ناشناخته برای ماشین تعریف می‌شوند. از جمله این مسائل می‌توان به مسائل سری زمانی و پردازش متن اشاره کرد. شبکه‌های بازگشتی شبکه‌هایی هستند که بعد از شبکه‌های پیشرو پیشنهاد شده و برای پردازش داده‌های متوالی مناسب می‌باشند. بسیاری از شبکه‌های بازگشتی از جمله شبکه‌های هاپفیلد

[33] برای بررسی داده‌های پیوسته و متوالی مناسب نیستند اما در دیگر مسائل عملکرد موفق داشته‌اند. در این پژوهش از شبکه بازگشتی GRU استفاده شده است که در ادامه توضیح داده خواهد شد.

واحد بازگشتی درگاه[1] (GRU) معماری تغییریافته از مدل LSTM است که به ترکیب حالت‌های پنهان شبکه (h_t) با مقادیر ذخیره شده در سلول حافظه (c_t) و ترکیب ورودی و گیت‌های فراموشی در قالب گیت (z_t) با اعمال یکسری تغییرات کوچک دیگر ایجاد می‌شود [34]. نمای کلی از شبکه GRU برمبنای روابط ۹-۶ می‌باشد:

$$r_t = \sigma(W_{xr}x_t + W_{hr}h_{t-1} + b_r) \quad (۶)$$

$$Z_t = \sigma(W_{xz}x_t + W_{hz}h_{t-1} + b_z) \quad (۷)$$

$$\tilde{h}_t = tanh(W_{xh}x_t + W_{hh}(h_{t-1} \odot r_t) + b_h \quad (۸)$$

$$h_t = (1 - Z_t) \odot \tilde{h}_t + Z_t \odot h_{t-1} \quad (۹)$$

مشابه LSTM عملگر ⊙ ضرب نقطه‌به‌نقطه بین بردارها اشاره می‌کند. GRU از ۴ ماتریس وزن، ۳ بایاس و یک متغیر حالت، تشکیل شده است. در پژوهش‌های صورت‌گرفته نتایج حاصل از GRU نشان می‌دهد که این مدل کارایی مشابه و هم‌سطح و در برخی موارد عملکرد بهتری در مقایسه با مدل LSTM دارا می‌باشد. در [34] گونه‌های متفاوتی از مدل‌های بازگشتی برروی مجموعه دادگان موسیقی پلیفونیک[2] اعمال شده و نتایج با یکدیگر مقایسه شده است. نتایج نشان داده است که GRU و LSTM به‌طور قابل توجهی عملکرد بسیار بهتری نسبت به دیگر معماری‌های بازگشتی داشته‌اند. در رویکرد ارائه‌شده در [35] نشان داده شده است که میانگین تغییرات GRU کمی بهتر از LSTM بوده است. در [36] رویکردی ارائه شده است که در آن برروی دادگان مختلف این دو معماری تست شده و نشان داده شده است که در اغلب موارد به استثنای عملیات مدلسازی زبان، GRU عملکرد بهتری نسبت به LSTM داشته است.

به‌طورکلی محققان بر این باورند که اکثر مدل‌های تغییریافته LSTM، از جمله GRU، تقریبا در همان سطح عملکرد و یا حتی در سطح بالاتری نسبت به LSTM هستند. همانطور که بیان شد تغییرات معرفی شده در GRU منجر به سادگی بیشتر معماری آن نسبت به مدل استاندارد LSTM شده است اما با توجه به اینکه سطح عملکرد این مدل‌ها مشابه می‌باشند GRU به‌طور فزاینده‌ای محبوب شده است. لذا در رویکرد پیشنهادی این مقاله از شبکه بازگشتی GRU در بخش رمزگشا استفاده شده است.

[1] Gated Recurrent Unit
[2] Polyphonic

۴- توصیف محتوای تصویر

با رشد روزافزون تصاویر دیجیتال، روزانه منجر به رشد منابع عکس مختلف از قبیل اینترنت، مقالات خبری و تبلیغات شده است. این منابع شامل تصاویری است که تماشاگران باید خودشان آنها را تفسیر کنند. اکثر تصاویر، توصیفی ندارند، اما انسان‌ها می‌توانند بدون توضیحات دقیق خودشان آنها را درک کنند. با این حال، درصورتی که انسان نیاز به توصیف خودکار تصاویر توسط ماشین داشته باشد، ماشین نیاز به تفسیر بسیاری از توصیف‌ها برای تصاویر دارد.

توصیف تصویر، یک موضوع تحقیقاتی محبوب در حوزه هوش مصنوعی است که با درک تصویر و استفاده از پردازش زبان طبیعی، به توضیح محتوای آن تصویر می پردازد. فهم تصویر نیاز به شناسایی و تشخیص اشیای درون تصویر، درک صحنه، خواص شی و ارتباط بین اشیا دارد. تولید جملات به‌منظور توصیف تصاویر نیز نیاز به درک قواعد نحوی و معنایی زبان دارد [37].

درک تصویر تاحد زیادی به ویژگی‌های محاسبه‌شده از تصاویر بستگی دارد. روش‌های مورد استفاده برای این منظور می‌توانند به‌طور گسترده‌ای به دو دسته تقسیم شوند: (۱) روش‌های مبتنی بر یادگیری ماشین سنتی[3] و (۲) روش های مبتنی بر یادگیری ماشین عمیق[4].

در روش‌های مبتنی بر یادگیری ماشین سنتی ، ویژگی‌های تصاویر به‌صورت دستی توسط کاربر محاسبه می‌شوند. در این روش‌ها از ویژگی‌هایی همچون الگوی دودویی محلی[5] (LBP) [38]، تبدیل ویژگی مقیاس نابسته[6] (SIFT) [39]، هیستوگرام گرادیان‌گرا[7] (HOG) [40] و ترکیبی از چنین ویژگی‌هایی به‌طور گسترده استفاده می‌شود.در این روش‌ها ویژگی‌ها از داده‌های ورودی استخراج می‌شوند و سپس این ویژگی‌ها به‌منظور طبقه‌بندی اشیای درون تصاویر، به یک طبقه‌بند مانند ماشین بردار پشتیبان[8] (SVM) [41] منتقل می‌شوند. به‌طورکلی ویژگی‌هایی که به‌طور دستی استخراج می‌شوند برای امور خاص مورد استفاده قرار می‌گیرند. همچنین استخراج ویژگی ازمجموعه تصاویر متنوع امکان پذیر نیست. علاوه‌براین، داده‌های دنیای واقعی، مانند تصاویر و ویدئوها پیچیده هستند و تفسیرهای معنایی متفاوتی می‌توان از آنها داشت.

از سوی دیگر، در روش‌های مبتنی بر یادگیری عمیق، ویژگی‌ها به‌طور خودکار از داده‌های آموزشی آموخته می‌شوند و می‌توانند

[3] Traditional machine learning based techniques
[4] Deep machine learning based techniques
[5] Local Binary Patterns
[6] Scale-Invariant Feature Transform
[7] Histogram of Oriented Gradients
[8] Support Vector Machine

مجموعه‌ای وسیع و متنوعی از تصاویر و فیلم‌ها را دربربگیرند. به‌عنوان مثال، شبکه‌های CNN برای یادگیری ویژگی‌ها به‌طور گسترده‌ای برای مسائل طبقه‌بندی مورد استفاده قرار می‌گیرند. همچنین می‌توان از ترکیب این شبکه‌ها با شبکه‌های بازگشتی به تولید و ایجاد توصیف برای تصاویر استفاده کرد.

در سال‌های اخیر، تعداد زیادی از مقالات بر روی اسناد تصویری با استفاده از آموزش عمیق ماشین مورد استفاده قرار گرفته‌اند. الگوریتم‌های یادگیری عمیق قادر به کنترل پیچیدگی و چالش‌های فرآیند توصیف تصاویر می‌باشند. ازجمله مقالاتی که به بازنگری مقالات منتشرشده در حوزه فرآیند توصیف تصویر می‌پردازند می‌توان به موارد [42], [12], [11] اشاره کرد.

۵- بیان مسئله

فرآیند توصیف تصاویر، از جمله روش‌های یادگیری ماشین است که قادر به توصیف محتوای تصاویر می‌باشد. عملیات صورت‌گرفته در این روش‌ها تنها در دسته‌بندی یا تشخیص اشیای درون تصویر خلاصه نمی‌شود. به‌طورکلی فهم تصویر نیاز به شناسایی و تشخیص اشیای درون تصویر، درک صحنه، خواص شی و تعاملات آنها دارد. در این مدل‌ها تشخیص و بکارگیری وابستگی‌های میان اشیای درون تصویر و ویژگی‌های متعلق به آنها و همچنین کدگذاری صحیح این اطلاعات در قالب توصیف تصاویر با بهره‌گیری از فرآیند پردازش زبان طبیعی، امری مهم تلقی می‌شود. با رشد روزافزون اطلاعات متنی و تصویری در اینترنت، وجود توصیف‌های معنادار برای تصاویر، نقش کلیدی در حوزه جستجو و بازیابی تصاویر مرتبط، ایفا می‌کنند.

اگر چه در سال‌های اخیر موفقیت‌های زیادی در این زمینه به دست آمده است، اما هنوز فرصت بزرگی برای بهبود کیفیت توصیف‌های ارائه‌شده برای تصاویر وجود دارد. روش‌های توصیف‌تصویر موجود، قادر به تولید متن‌های جدید برای هر تصویر هستند. بااین‌حال، این روش‌ها در تولید عبارات دقیق و چندگانه برای تصاویر، قادر به شناسایی اشیا برجسته تصویر، ویژگی‌ها و روابط میان آنها به میزان قابل توجهی نیستند. علاوه‌براین، دقت توصیف‌های تولید شده به‌طور عمده به متن‌های صحیح و متنوعی بستگی دارد که مدل تولید زبان در فاز یادگیری به آنها متکی است.

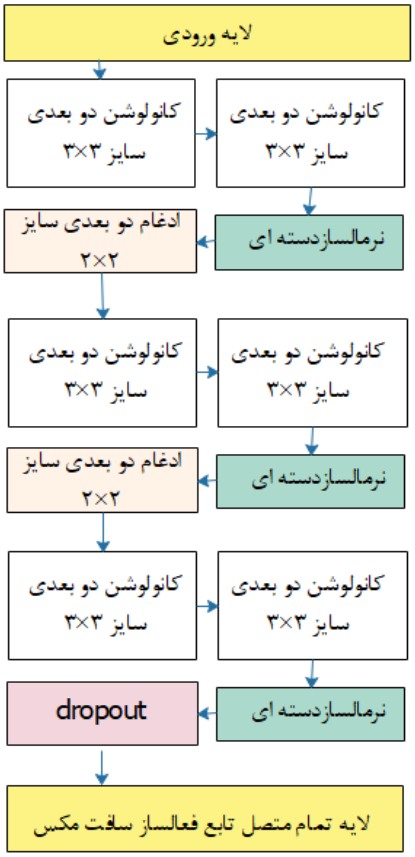
ارائه روابط انتزاعی بین اشیای درون تصویر در مدل‌های یادگیری ماشین بسیار چالش‌برانگیز است. بنابراین باید به دنبال روش‌های ارزشمندی باشیم که بتواند این روابط انتزاعی را فراهم کند. بسیاری از روش‌های کنونی برای تولید توضیحات برای تصاویر نتوانسته‌اند به طور قابل‌توجهی اشیا برجسته درون تصویر، ویژگی‌ها و روابط بین آنها را شناسایی کنند. دقت توصیفات تولید شده به دقت و تنوع متنی بستگی دارد که مدل تولید زبان در مرحله یادگیری بر آن تکیه دارد.

همانطور که در بخش مرور پژوهش‌های اخیر بیان شد، بسیاری از مدل‌های پیشنهادی عنوان تصویر فعلی بطور عمده از CNNاستفاده می‌کنند. CNNها نیز دارای چند اشکال و محدودیت هستند. این شبکه‌ها اطلاعات مکانی (مانند موقعیت شی، سایز و جهت) اشیای درون تصویر و روابط فضایی نسبی بین ویژگی‌ها را در حین انجام پیش‌بینی‌ها را درنظر نمی‌گیرند. برای مثال اگر تصویری از صورت انسان با اجزای درهم (مثلا جای بینی با لب عوض شود) به هر دو شبکه کپسولی و CNN بدهیم به‌عنوان خروجی شبکه CNN با درصد بالایی از احتمال، عکس را به عنوان صورت انسان تشخیص می‌دهد زیرا صرفا به اجزای صورت به‌طور متمایز نگاه می‌کند و موقعیت اجزا نسبت به یکدیگر را درنظر نمی‌گیرد درصورتی‌که این احتمال برای شبکه کپسول پایین است زیرا موقعیت لب با بینی و نسبت به دیگر اجزای صورت در این شبکه در نظر گرفته می‌شود. یکی دیگر از اشکال‌های اصلی CNNها این است که این شبکه‌ها فاقد ویژگی‌های تغییرناپذیری و ثبات در برابر چرخش[1] هستند. به نظر می‌رسد آنها به راحتی پیش‌بینی خود را به اشتباه در هنگام چرخش جسم تغییر می‌دهند. برای غلبه بر اشکال‌ها و محدودیت های ذکر شده، صبور و همکارانش در [13] شبکه کپسولی را به‌عنوان جایگزین قدرتمندی برای CNN معرفی کردند. در CapsNet، خروجی شبکه به‌جای یک مقدار عددی، به‌صورت بردار بوده که طول بردار احتمال وجود یک موجودیت در تصویر و جهت بردار نشان‌دهنده ویژگی‌های موجودیت است. CapsNet همچنین عملیات ادغام در CNN را با مفهوم مکانیزم مسیریابی پویا جایگزین می‌کند که بر نواحی برجسته تصویر تمرکز داشته و به استخراج ویژگی از آن نواحی می‌پردازد.

در این مقاله از یک شبکه کپسولی با ساختار بهبودیافته، به‌عنوان استخراج‌کننده ویژگی‌های بصری از تصاویر استفاده می‌کنیم. ویژگی‌های استخراج شده را به همراه اطلاعات متنی تصاویر به عنوان ورودی به شبکه RNN می‌دهیم. در شبکه بازگشتی به‌منظور نگاشت بردار کلمات موجود در توصیف تصاویر به یک دنباله متناظر از اعداد حقیقی از یک روش تعبیه‌سازی کلمه به‌نام ELMo[2] استفاده می‌شود. برای استخراج ویژگی از تصاویر از یک شبکه پیچشی با شش لایه پیچشی، سه لایه نرمال‌سازی دسته‌ای، دو لایه ادغام که به ترتیب پس از دو لایه اول نرمال‌سازی دسته‌ای قرار گرفته‌اند، استفاده می‌شود. در انتهای شبکه سه لایه تمام متصل به طول ۲۰۴۸ وجود دارند که به یک لایه سافت‌مکس به طول ۱۰۰۰ متصل می‌گردند. به‌منظور جلوگیری از بیش‌برازش شبکه برروی دادگان آموزشی از یک لایه dropout بعد از سومین لایه نرمال‌ساز دسته‌ای استفاده می‌کنیم. لایه dropout تنها در فاز

[1] Rotation Invariance
[2] Embeddings from Language Model

آموزش مورد استفاده قرار می‌گیرد. شکل ۳ نمایی از این شبکه را نمایش می‌دهد.



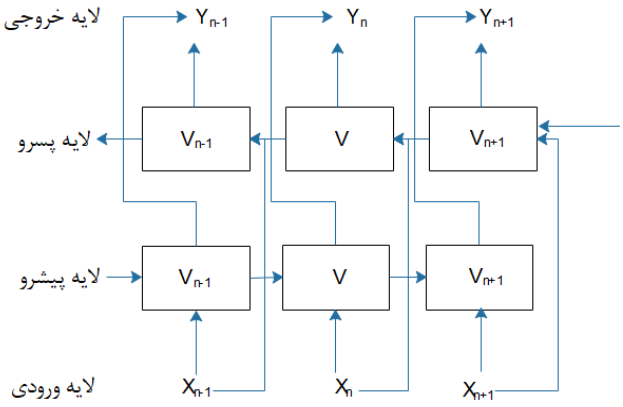
شکل (۳) ساختار شبکه پیچشی پیشنهادشده در مقاله

برای ارزیابی نتایج، از برخی معیارهای گسسته در پردازش زبان طبیعی مانند BLEU 1-4 [43]، ROUGE [44] و METEOR [45] برای بررسی عملکرد مدل پیشنهادی استفاده می‌کنیم. در بخش شش به طور مختصر این معیارها را توضیح خواهیم داد.

ELMo -۱-۵

روش تعبیه‌سازی کلمه فرایندی است که متن را به نمودی ساده و قابل فهم برای ماشین تبدیل می‌کند. روش‌های متفاوتی برای تعبیه-سازی کلمه‌های موجود در متن وجود دارد که از جمله آنها می‌توان به روش های Word2Vec، GloVe و ELMO اشاره کرد. این بردار از اعداد منحصربفرد ارائه می‌شوند که هر عدد نمایشگر یک لغت از مجموعه لغات درنظر گرفته شده می‌باشد. با نمایش کلمات موجود در متن به‌صورت بردار امکان انجام عملیات ریاضی و اعمال برخی معادلات جبری خطی برای اندازه‌گیری فاصله میان بردارها و بررسی ارتباط میان واژگان فراهم می‌شود.

در این مقاله ما از روش تعبیه‌سازی واژگان ELMo که توسط پیتر و همکارانش [46] ارائه شده است برای یادگیری توضیحات متنی تصاویر استفاده می‌کنیم. ELMo یک روش نوین برای تعبیه‌سازی کلمات است که مدل‌های پردازش زبان طبیعی کاربرد دارد. این روش به مدل‌سازی ویژگی‌های پیچیده زبانی مانند نحو و معنای جملات وچگونگی تغییر این کاربردها در متون زبانی متنوع است (به عنوان مثال برای مدل‌سازی کلمات با چندمعنی) می‌پردازد. این بردارهای کلمه، توابع آموخته شده از حالات داخلی یک مدل زبان دوطرفه عمیق (biLM) هستند که از قبل بر روی یک مجموعه متن بزرگ آموزش داده شده است و پارامترهای یادگیری شده برای استفاده از کاربردهای مختلف در مدل به اصلاح فریز می‌شوند. هر لایه از این شبکه یک نمایش برداری را برای هر کلمه از جمله محاسبه می‌کند. شکل 4 معماری این مدل را نمایش می‌دهد.



شکل (۴) ساختار داخلی شبکه ELMo

همان‌طور که مشهود است، در این مدل هرلایه شامل اتصالات پیشرو و پسرو می‌باشد. در فاز پیشرو یک مدل زبانی LSTM بر روی دنباله کلمات در جمله از ابتدا تا انتها اعمال می‌شود به این صورت که در هر مرحله مدل سعی می‌کند تا کلمه t_k را بر مبنای کلمه‌های پیشین $t_1, t_2, \dots, t_{k-1}$ یا به اصطلاح پیشوندهای آن پیش‌بینی کند. X_n یک نمایش منحصربفرد برای هر کلمه می‌باشد. Y_n خروجی حاصل از هر شبکه LSTM و نشان دهنده بردار جاسازی شده متن ورودی است.

$$p(t_k \mid t_1, t_2, \dots, t_{k-1}; \Theta_x, \vec{\Theta}_{LSTM}, \Theta_s) \qquad (۱۰)$$

در این رابطه t_k، k اُمین کلمه مورد بررسی از یک جمله حاوی N کلمه است. Θ_x بردار اولیه تعبیه‌سازی شده از کلمه، $\vec{\Theta}_{LSTM}$ بردارحاصل از مدلسازی زبانی توسط شبکه LSTM و Θ_s خروجی لایه سافت‌مکس بر مبنای برچسب‌های دادگان آموزشی می‌باشد. سافت‌مکس یک توزیع احتمالی بر روی لغات ایجاد می‌کند که برای پیش‌بینی کلمه بعدی در هر مرحله مورد استفاده قرار می‌گیرد.

در فاز پسرو نیز عملیاتی مشابه اما در جهت عکس انجام می‌گیرد. در این فاز، مدل LSTM بر روی دنباله کلمات در جمله از انتها به ابتدا اعمال می‌شود به این صورت که در هر مرحله مدل سعی می-کند تا کلمه t_k را بر مبنای کلمه‌های پسین $t_{k+1}, t_{k+2}, \dots, t_N$ یا به اصطلاح پسوندهای آن، پیش‌بینی کند. در نهایت مدل زبانی به-

طور مشترک با ترکیب روابط فوق با استفاده از رابطه ۱۱ آموزش داده می‌شود.

$$p(t_k \mid t_{k+1}, t_{k+2}, \ldots, t_N;\ \theta_x, \tilde{\theta}_{LSTM}, \theta_s) \quad (۱۱)$$

درنهایت یک تابع برروی ورودی اعمال می‌شود که این تابع مجموع k تابع خطی متمایز را به‌عنوان ورودی می‌گیرد و احتمال عضویت در j اُمین کلاس موجود برای بردار ورودی x و بردار وزن w را با استفاده از رابطه 12 محاسبه می‌کند.

$$P(y=j|X) = \frac{e^{X^T W_j}}{\sum_{k=1}^{K} e^{X^T W_k}} \quad (۱۲)$$

برخلاف Glove و Word2Vec، ELMo به تعبیه‌سازی یک کلمه با استفاده از جمله کامل حاوی آن کلمه نشان می‌دهد. بنابراین، تعبیه‌های ELMo می‌توانند معنای کلمه مورد استفاده در جمله را درنظربگیرد. این روش قادر است با توجه به معنای کلمه بر مبنی سایر کلمات درون جمله تعبیه‌سازی‌های مختلفی را برای همان کلمه در جملات مختلف ایجاد کند. از ELMo می‌توان به‌عنوان لایه ورودی به یک شبکه مدل‌سازی زبان استفاده کرد. از این‌رو در این مقاله نیز از روش تعبیه‌سازی ELMo به‌عنوان یک لایه ورودی به شبکه مدل‌سازی زبان با سه لایه GRU استفاده شده است که این شبکه به پیش‌بینی توصیف محتوای تصویر می‌پردازد.

۵-۲- چارچوب مدل پیشنهادی

مدل نهایی از چارچوب رمزگذار-رمزگشا پیروی می‌کند. معماری مدل پیشنهادی در شکل ۵ نشان داده شده است. در این مدل دو فاز اصلی وجود دارد. فاز اول شامل استخراج ویژگی‌ها از تصاویر با استفاده از دو شبکه عصبی عمیق است. در این مرحله از شبکه کپسولی و یک شبکه پیچشی برای استخراج محتوای بصری از تصویر ورودی به‌طور همزمان استفاده می‌شود.

در شبکه کپسولی ابتدا سه لایه پیچشی با اندازه‌های ۷۲×۷۲×۹۶، ۳۴×۳۴×۹۶ و ۲۶×۲۶×۲۵۶ روی تصویر اعمال می‌شود (شکل ۵.الف). همانطور که در بخش 4.2 بیان شد، یک لایه کپسول اولیه با یک فرآیند تغییر شکل[1] و اسکواشینگ[2] دنبال می‌شود تا ویژگی‌های اساسی شناسایی شده توسط لایه‌های پیچشی را بگیرد و آنها را برای تولید ویژگی‌های جدید ترکیب کند (شکل ۵.ب). سپس مکانیزم "مسیریابی با توافق"[3] به جای عملیات ادغام انجام می‌شود. بر اساس این مکانیزم، خروجی هر کپسول در سطح پایین‌تر با وابستگی به ویژگی‌های آن‌ها به کپسول‌های مادر در سطح بالاتر ارسال می‌شود. لایه بعدی کپسول‌های طبقه‌بندی است که نشان‌دهنده احتمال عضویت تصویر ورودی در هر دسته است. برچسب واقعی برای پوشاندن خروجی لایه کپسول طبقه‌بندی با استفاده از نرم L_2 برای محاسبه میزان خطای شبکه استفاده می‌شود (شکل ۵.پ). آخرین قسمت شبکه کپسولی رمزگشا است که به‌عنوان یک تنظیم‌کننده با دو لایه کاملاً متصل با اندازه‌های ۵۱۲ و ۱۰۲۴ استفاده می‌شود (شکل ۵.ت). در فاز رمزگذاری کپسول‌ها مجبور به یادگیری ویژگی‌هایی هستند که می‌توان از آنها برای بازسازی تصویر اصلی توسط رمزگشا استفاده کرد. خروجی دومین لایه تمام متصل به عنوان بردار ویژگی‌های بصری تصویر استفاده می‌شود. (شکل ۵.ج).

شبکه پیچشی با شش لایه پیچشی، به‌عنوان دومین استخراج کننده ویژگی‌های بصری از تصویر ورودی می‌باشد (شکل ۵.د). در مرحله بعد، هر دو بردار ویژگی بصری استخراج شده از شبکه کپسولی و شبکه پیچشی با یکدیگر ادغام می‌شوند و به مدل زبانی وارد می‌شوند (شکل ۵.ذ).

در نهایت، اطلاعات از فاز اول به آخرین فاز تغذیه می‌شود. در آخرین مرحله چارچوب، از شبکه بازگشتی با سه لایه GRU به عنوان رمزگشا استفاده می‌کنیم (شکل ۵.ز).لایه‌های توکن‌سازی و تعبیه‌سازی ELMo، تمام داده‌های متنی از پیش‌پردازش شده را قبل از وارد کردن توضیحات به مدل زبان، به برداری از اعداد تبدیل می‌کنند. (شکل ۵.ر). درنهایت، مدل ما آموزش می‌دهد تا تمام ویژگی‌های متنی و بصری تصاویر را با استفاده از روش‌های مدل‌سازی زبان توصیف کند (شکل ۵.ژ).

پس از مرحله آموزش، مدل با استخراج ویژگی‌های بصری و پیش‌بینی زیرنویس‌ها با استفاده از جستجوی حریصانه، آماده ارزیابی تصاویر مجموعه آزمایشی است. جستجوی حریصانه، کلمه‌ای را با بیشترین احتمال در هر مرحله زمانی انتخاب می‌کند و از آن به‌عنوان ورودی GRU برای مرحله زمانی بعدی استفاده می‌کند تا به پایان جمله برسد. در بخش بعدی جزئیات آزمایش‌ها و نتایج به‌دست آمده با روش‌های تحلیل شده بررسی می‌شود.

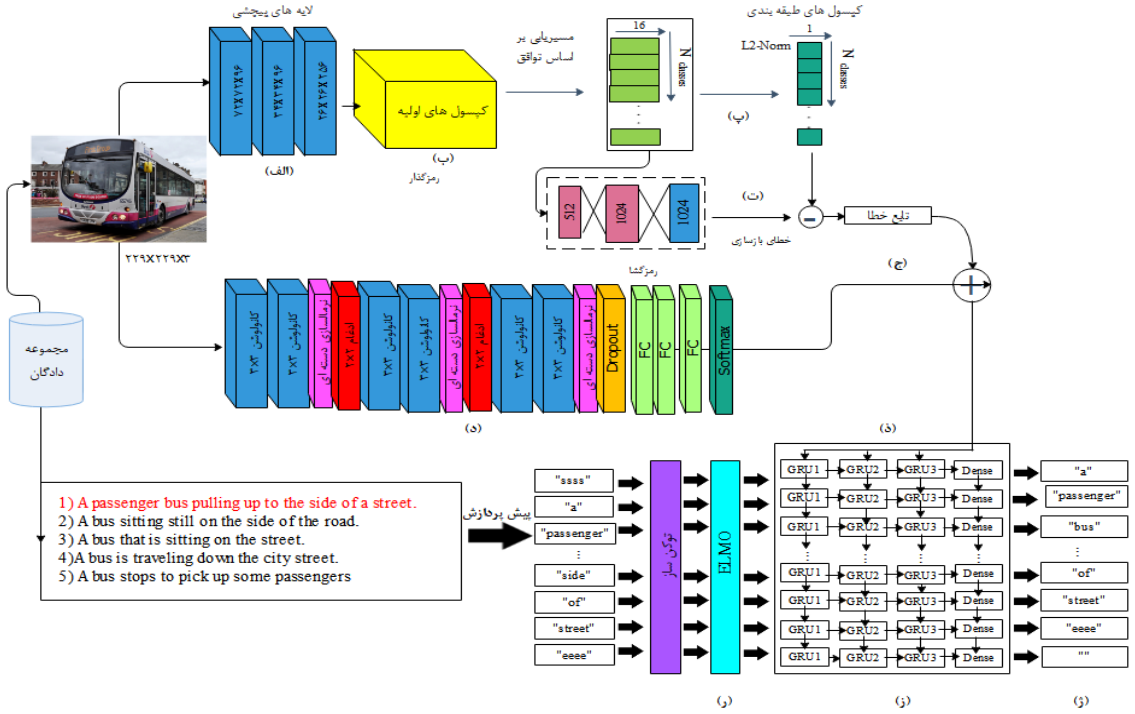
۶- پیاده‌سازی و نتایج

در این بخش جزئیات پیاده‌سازی‌ها و نتایج آزمایش‌های انجام شده توسط انواع مختلف مدل‌ها گزارش می‌شود.

۶-۱- مجموعه داده و معیارهای ارزیابی

در این مقاله از مجموعه داده MS-COCO [47] برای ارزیابی مدل پیشنهادی در آزمایش‌های ارائه شده استفاده شده است. MS-COCO شامل ۱۲۳۲۸۷ تصویر با پنج عنوان و ۸۰ دسته شی برای هر تصویر است که توسط کارگران آمازون (AMT)[4] حاشیه نویسی شده است. از آنجایی که هیچ حاشیه نویسی در دسترس برای مجموعه آزمایشی وجود ندارد، در این مقاله، از تقسیم‌بندی‌های ارائه شده توسط Karpathy و همکاران استفاده شده است [48].

[1] Reshaping
[2] Squashing
[3] Routing by Agreement
[4] Amazon Mechanical Turk (AMT)



شکل ۵ چارچوب مدل پیشنهادی

از مجموعه دادگان مورد بررسی ۵۰۰۰ تصویر برای اعتبارسنجی و تست و بقیه برای مجموعه آموزشی استفاده شده است.همه مدل‌ها در پایتون نسخه ۳,۶ و با استفاده از کتابخانه Keras نسخه ۲,۲,۵ و TensorFlow نسخه ۱,۱۵,۰ پیاده‌سازی شده است. جدول ۱ پارامترهای تنظیم‌شده برای هر شبکه را نشان می‌دهد.

جدول 1 مقدار دهی اولیه پارامترهای ارزیابی شده

پارامترهای یادگیری	شبکه های یادگیری عمیق	
	Convolutional	CapsNet
Depth	۱۲	6
Image size (pixel)	229×229×3	229×229×3
Solver (optimizer)	RMSProb	ADAM
Loss function	cross entropy	MSE
Batch size	64	128
Learning rate	0001/0	001/0
Gradient threshold method	L_2norm	L_2norm

کلیه پیاده‌سازی‌ها با استفاده از دستگاه مجهز به دو کارت گرافیک GeForce RTX 2080 با حافظه ۸ گیگابایتی انجام شده است.

معیارهای ارزیابی: به‌منظور مقایسه نتایج خود با سایر مدل‌های پایه، عملکرد مدل‌های پیاده‌سازی شده را با معیارهای رایج مورد استفاده در حوزه توصیف تصاویر اندازه‌گیری می‌کنیم. این معیارها عبارتنداز: BLEU 1-4 [43] ، ROUGE [44] و METEOR [19].

BLEU-n یکی از معیارهای محبوب برای ارزیابی مطابقت بین جملات تولید شده توسط انسان و ماشین است. این معیار حداکثر تعداد n -گرم های مختلف همزمان بین جمله‌های مرجع و نامزد را اندازه‌گیری می‌کند. در اینجا 'n' مقدار ۱، ۲، ۳، ۴ را بسته به طول جمله می‌گیرد. هر معیار BLEU-N میانگین دقت‌های محاسبه شده از n=1 تا n=N می‌باشد. این معیار حداکثر تعداد n-گرم‌های مختلف همزمان بین جمله‌های مرجع و نامزد را اندازه‌گیری می‌کند. به این معنی که BLEU-1 دقت توصیف ایجاد شده برای تصویر با توصیف مرجع بر مبنای ۱-گرم، BLEU-2 میانگین هندسی دقت های محاسبه شده بر مبنای ۱-گرم و ۲-گرم، BLEU-3 میانگین هندسی دقت‌های محاسبه شده بر مبنای ۱-گرم و ۲-گرم و ۳-گرم، و در نهایت BLEU-4 میانگین هندسی دقت های محاسبه شده بر مبنای 1-گرم و ۲-گرم و ۳-گرم و ۴-گرم می‌باشد.

ROUGE عملکرد جملات تولید شده توسط یک ماشین را بر اساس شباهت آنها به جملات مرجع ارزیابی می‌کند. این معیار طولانی‌ترین دنباله توکن‌ها را بین جملات کاندید و مرجع پیدا و محاسبه می‌کند که چه تعداد نشانه از خلاصه‌های مرجع انسانی در خلاصه‌های تولید شده توسط ماشین کپی شده‌اند. برخلاف BLEU که دقت را در اولویت قرار می‌دهد، ROUG فراخوانی-گرا است و می‌تواند n-گرم‌های مرتبط را بهتر از BLEU تخمین بزند.

METEOR آخرین معیار ارزیابی در این مقاله است. این معیار، هم‌ترازی کلمات در جملات کاندید و مرجع را با درنظرگرفتن ریشه کلمات یا مترادف آنها در پایگاه داده وردنت[1] بررسی می‌کند.

مدل پایه: در این مقاله یک مدل پایه برای تأیید اثربخشی دیگر مدل‌های پیاده‌سازی شده درنظر گرفته شده است. چارچوب خط پایه تقریباً مشابه مدل [49] به‌عنوان یک روش پایه است با این تفاوت که GRU جایگزین مدل زبان LSTM می‌شود. در مدل پایه از inception-V3 به عنوان روش استخراج ویژگی برای بخش رمزگذار استفاده شده است.

۶-۲- نتایج و بحث

در این بخش به ارزیابی رویکرهای ارائه شده در این مقاله می‌پردازیم. مدل‌های مورد استفاده برای شرح‌گذاری تصاویر در این مقاله عبارتند از: CapsNet+CNN ، CapsNet+CNN+ELMo و CapsNet+ELMo. به‌منظور بررسی میزان کارآمدی مدل‌های ارائه شده در فرآیند توصیف تصاویر کلیه نتایج با نتایج حاصل از مدل‌های پایه مورد مقایسه قرار گرفته است.

به‌منظور محاسبه میزان عملکرد استفاده از روش تعبیه‌سازی ELMo، مدلی بدون استفاده از این روش تعبیه‌سازی و با استفاده از روش تعبیه‌سازی ساده Word2Vec طراحی و اجرا شد. در این مدل در بخش رمزگشا از روش Word2Vec به منظور تعبیه‌سازی کلمات استفاده شده است. Word2Vec یکی از محبوب‌ترین تعبیه‌سازی کلمات با استفاده از شبکه عصبی کم‌عمق است که یک بازنمایی متراکم از کلمات موجود در هر جمله در قالب بردارهای عددی ارائه می‌دهد. این توسط توماس میکولوف در سال ۲۰۱۳ در گوگل توسعه یافته است[50]. در این مدل، تعبیه‌سازی کلمات را می‌توان با استفاده از دو روش (هر دو شامل شبکه های عصبی) اسکیپ-گرام[2] و CBOW[3] بدست آورد. با توجه به استفاده از روش CBOW برای مجموعه دادگان کوچک، با توجه به مقیاس بالای دادگان مورد استفاده، در این مقاله از روش اسکیپ-گرام استفاده شده است. در ادامه به توضیح مختصر این روش ارائه شده و به‌منظور جلوگیری از پیچیدگی، از ارائه روابط ریاضی صرف‌نظر می‌شود. در روش اسکیپ-گرام یک پنجره متحرک با اندازه ثابت در نظر می‌گیریم و آن را در امتداد یک جمله حرکت می‌دهیم.

در این روش کلمه‌ی وسط به‌عنوان "هدف" در نظر گرفته می‌شود و کلماتی که در سمت چپ و راست این کلمه هدف در داخل پنجره متحرک قرار می‌گیرند همان کلمه‌هایی که برمبنای آنها کلمه هدف پیش‌بینی خواهند شد.

مدل مورد بررسی بعد مدل CapsNet+CNN+ELMo می‌باشد که تفاوت این مدل با مدل قبل در روش تعبیه‌سازی کلمات است که در این مدل از روش ELMo برای تولید بردارهای تعبیه‌سازی کلمه‌های درون متن مورد بررسی استفاده شده است.

آخرین مدل بررسی در این مقاله مدل CapsNet+ELMo می‌باشد که این مدل تنها از شبکه کپسولی در فاز رمزگذاری برای تولید ویژگی‌های تصویر استفاده می‌کند. در این روش نیز از روش ELMo برای تعبیه‌سازی و تولید بردارهایی با اعداد حقیقی از کلمات استفاده شده است.

نتایج حاصل از پیاده‌سازی مدل‌های معرفی شده برروی دادگان MS-COCO در جدول ۲ ارائه شده است. برای اختصار در توضیح نتایج، معیارهای1-4BLEU، METEOR و ROUGE را به ترتیب B1-B4، M و R می‌نامیم. در روش CapsNet+CNN+ELMo نتایج ارائه شده به‌ترتیب ۰/۸۳، ۰/۶۳، ۰/۵۵، ۰/۴۵، ۰/۵۶ و۰/۳۱می‌باشد که در همه معیارها به‌جز در معیار B2 و R این مدل بهترین عملکرد را از خود نشان داده است. نتایج ارائه شده برای این مدل در مقایسه با مدل پایه برای معیارهای ارزیابی B1-B4، M و R به‌ترتیب ۰/۲۴، ۰/۱۷، ۰/۲، ۰/۱۹، ۰/۱۴ و ۰/۱ بهبود داشته است.

جدول 2 نتایج مدل‌های ارزیابی شده و مقایسه با رویکردهای اخیر

Models	Metrics					
	B1	B2	B3	B4	R	M
CNN (Baseline)	0/59	0/46	0/35	0/26	0/42	0/21
CapsNet + CNN	0/72	0/53	0/46	0/37	0/53	0/27
CapsNet + CNN +ELMo	**0/83**	0/63	**0/55**	**0/45**	0/56	**0/31**
CapsNet + ELMo	0/68	0/49	0/41	0/34	0/47	0/25
Aneja et.al 2018 [51]	0/72	0/55	0/40	0/30	0/53	0/25
Tan et.al 2019 [52]	0/73	0/57	0/43	0/33	0/54	0/25
Wu et.al 2017 [49]	0/73	0/56	0/41	0/31	0/53	0/25
Zhang et.al 2021 [53]	0/75	0/62	0/48	0/36	-	0/27
Yu et.al 2019 [54]	0/81	**0/67**	0/52	0/40	**0/59**	0/29
Li et.al 2021 [55]	0/81	-	-	0/39	0/58	0/28
Anderson et.al 2018 [56]	0/80	0/64	0/49	0/37	0/57	0/27
Jiang et.al 2018 [57]	0/80	0/65	0/50	0/38	0/58	0/28
Yan et.al 2020 [58]	0/73	0/53	0/39	0/28	0/56	0/25

در مدل CapsNet+CNN، از دو شبکه کپسولی و پیچشی که در بخش ۴ معرفی شدند به‌عنوان استخراج‌کننده ویژگی استفاده می‌شوند. نتایج ارائه‌شده برای این مدل با نزول مقادیر ۰/۱۱، ۰/۱، ۰/۹، ۰/۰۸، ۰/۰۳ و ۰/۰۴ به ترتیب برای معیارهای ارزیابی B1-B4، M و R همراه بود. این در حالیست که نتایج در مقایسه با مدل پایه ۰/۱۳، ۰/۰۷، ۰/۱۱، ۰/۱۱، ۰/۱۱ و ۰/۰۶ به

[1] Wordnet
[2] Skip Gram
[3] Common Bag Of Words

ترتیب برای معیارهای ارزیابی ذکرشده بهبود داشت. نتایج ارائه شده برای این مدل نشان از عملکرد مناسب شبکه کپسولی در فرآیند توصیف تصاویر می‌باشد.

در نهایت کلیه آزمایش‌ها با استفاده از مدل کپسولی و روش تعبیه‌سازی ELMo و بدون مشارکت شبکه پیچشی تکرار و ارزیابی گردید. نتایج ارائه شده نشان می‌دهد که نتایج نسبت به مدل CapsNet+CNN+ELMo با نزول معیارهای ارزیابی مواجه شد.

در نهایت مدل CapsNet+ELMo در معیارهای ارزیابی مورد بررسی B1-B4 ، M و R به ترتیب با کاهش مقادیر ۰/۱۵، ۰/۱۴، ۰/۱۴، ۰/۱۱، ۰/۹ و ۰/۶ همراه بود. این در حالیست که نتایج فوق در مقایسه با مدل پایه دارای میزان بهبود ۰/۰۹، ۰/۰۳، ۰/۰۶، ۰/۰۸، ۰/۰۵ و ۰/۰۴ می‌باشد که این بهبود نشان از عملکرد خوب شبکه کپسولی در ترکیب با روش تعبیه‌سازی ELMo می‌باشد.

شکل ۶، عملکرد همه مدل‌های معرفی‌شده روی دادگان MS-COCO برای معیارهای ارزیابی BLEU 1-4، METEOR و در ۱۰۰ دوره آموزشی نشان می‌دهد. مقایسه نتایج ارائه شده با نتایج حاصل از مدل‌های پایه نشان می‌دهد که عملکرد مدل CapsNet+CNN+ELMo که به استخراج ویژگی‌های تصویر را با استفاده از شبکه پیچشی و شبکه کپسولی می‌پردازد و همچنین از روش ELMo برای تولید بردارهای تعبیه‌سازی کلمه‌های درون متن استفاده می‌کند، به طور قابل‌توجهی بهتر از مدل‌های دیگر عمل می‌کند. نمودارهای ارائه شده نشان می‌دهند که اکثر مدل ها پس از ۶۰ دوره آموزش همگرا شده اند.

برای اثبات اثربخشی این مدل، نتیجه روش CapsNet+CNN+ELMo نتایج بدست آمده از مدل‌های ارائه‌شده در این مقاله با نتایج حاصل از تحقیقات اخیر مقایسه می‌کنیم. نتایج حاصل حاکی از عملکرد خوب رویکرد ارائه شده می‌باشد. جدول ۲ نشان می دهد که بهترین مدل ما از نتایج منتشر شده قبلی بر مجموعه داده MS-COCO با تقسیم‌بندی ارائه‌شده توسط "Karpathy" عملکرد خوب رویکرد ارائه داده است. در مقایسه با مدل ما، [51] مکانیزم توجهی را برای استفاده از ویژگی‌های فضایی یک تصویر برای یافتن اشیاء برجسته پیشنهاد کرده است. تان و همکاران در [52] یک مدل تنظیم با تعداد کمی از پارامترها در RNN پیشنهاد کرده‌اند. مدل آن‌ها می‌تواند رمزگشای بسیار پراکنده‌ای را برای ایجاد عنوانی تولید کند که عملکرد روش را در مقایسه با خط پایه خود حفظ می‌کند. ژانگ و همکاران در [53] یک مکانیزم یادگیری مشارکتی برای توصیف تصویر ایجاد کردند که این مکانیزم به ترکیب دو ماژول توصیف تصویر و بازیابی‌کننده تصویر می‌پردازد. سپس طی یک فرآیند پالایش چند مرحله‌ای، آنها اطلاعات سطح تصویر و سطح شی را برای ایجاد یک عنوان معنی‌دار اصلاح کردند.

یو و همکاران [54] مدلی بر مبنای مبدل‌ها جهت توصیف تصاویر پیشنهاد می‌دهند. در مدل پیشنهادی آنها، از مبدل‌های چندوجهی به عنوان مدل زبانی در شبکه بازگشتی در فاز رمزگشایی برای تولید توصیف متناسب با محتوای تصویر استفاده می‌شود. رویکرد ارائه‌شده توسط آنها در معیارهای ارزیابی BLEU2 و ROUGE با اندک اختلاف ۴ و ۳ درصد در مقایسه با بهترین رویکرد ارائه‌شده در این مقاله به ترتیب به نتایج بهتری دست یافته است.

در [56] اندرسون و همکاران با اعمال مکانیزم از پایین به بالا، مجموعه‌ای از مناطق برجسته را از تصویر استخراج می‌کنند. آنها همچنین یک مکانیزم از بالا به پایین برای تعیین توزیع توجه بر روی تصویر برای محاسبه وزن ویژگی ها در مناطق مختلف اجرا کردند.

جیانگ و همکاران [57] چارچوبی را پیشنهاد کردند که شامل یک شبکه همجوشی مکرر است. این روش ادغام بین رمزگذار و رمزگشا اجرا می‌شود تا از تعاملات بین ویژگی های نمایش داده شده از قسمت رمزگذار برای ایجاد مجموعه جدیدی از بردارها از خروجی های رمزگشا استفاده کند. در بخش بعد به ارائه نمونه‌های کیفی از نتایج حاصل از مدل‌های ارزیابی شده می‌پردازیم.

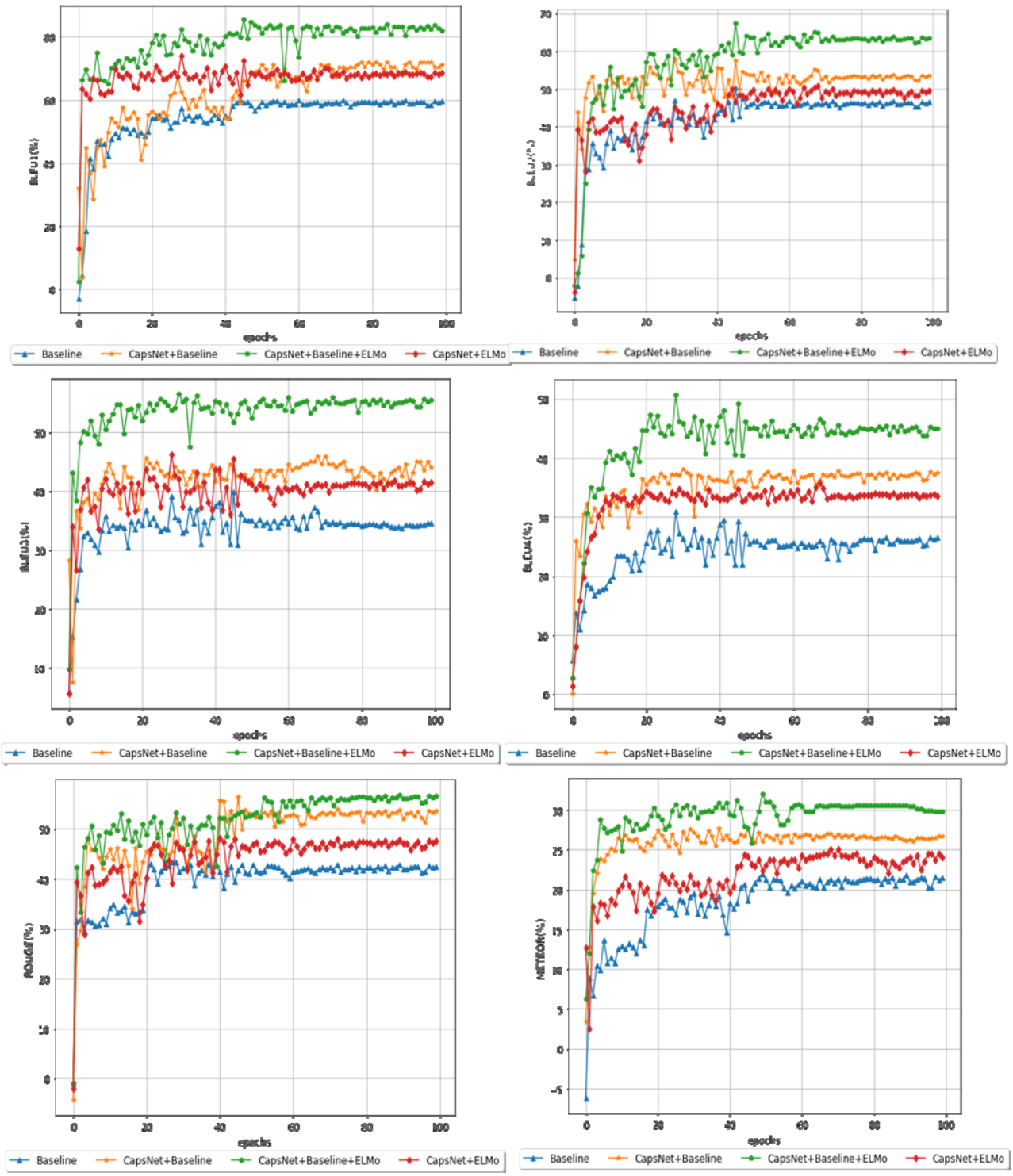
۶-۳- نتایج کیفی

در این مقاله، فرضیه اولیه ما تاثیر مثبت استفاده از شبکه کپسولی به‌منظور استخراج ویژگی‌های بامعناتر توسط شبکه در فرآیند تولید توصیف برای تصاویر بود که نتایج حاصل از آزمایش-های صورت گرفته نشان از موثر بودن این ایده می‌باشد. همچنین بهره‌مندی از روش تعبیه‌سازی ELMo در بخش رمزگشا باعث افزایش قابلیت مدل زبانی در تولید توصیف‌هایی با نحو و ساختار بهتر برای تصاویر شد. موثر بودن استفاده از شبکه کپسولی و روش تعبیه‌سازی ELMo در کیفیت توصیف‌های تولیدشده برای تصاویر با استفاده از پارامترهای ارزیابی شده در جدول ۲ در بخش قبل نشان داده شده است. نتایج ارائه‌شده نشان می‌دهد که در صورت عدم استفاده از شبکه کپسولی در مدل CNN (Baseline) دقت‌های ارزیابی‌شده در تمامی پارامترها بسیار کمتر از زمانیست که علاوه بر شبکه CNN، در مدل CapsNet + CNN از شبکه کپسولی نیز به‌عنوان استخراج‌کننده ویژگی استفاده شده است.

مقایسه نتایج حاصل از دو مدل CapsNet + CNN +ELMo و CapsNet + ELMo نشان می‌دهد که استفاده از روش تعبیه‌سازی ELMo منجر به بهبود نتایج ارائه‌شده می‌شود. لذا با توجه به نتایج حاصل فرضیه اولیه تحقیق بدین ترتیب به اثبات می‌رسد. شکل ۷، نمونه‌هایی را برای نشان دادن عملکرد مناسب روش CapsNet + CNN + ELMo به عنوان بهترین مدل این تحقیق ارائه می‌دهد.

در این شکل، از تابع حساسیت انسدادظ برای تجسم و محلی-سازی[1] مهم‌ترین مناطق تصاویر برای شبکه استفاده شده است. این تابع نواحی ورودی را با یک ماسک مسدودکننده[2] که معمولاً یک مربع خاکستری می‌باشد جایگزین می‌کند سپس ماسک را در

[1] Visualize and Localize
[2] Occluding Mask

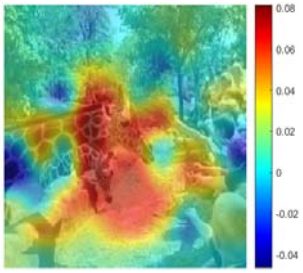
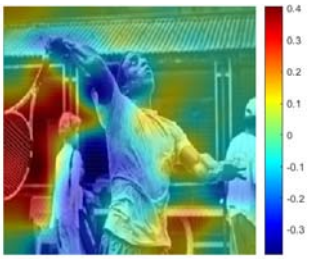
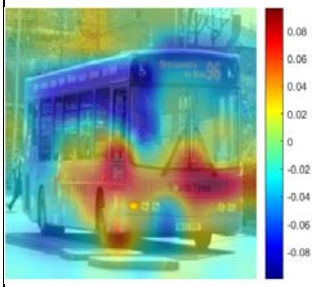
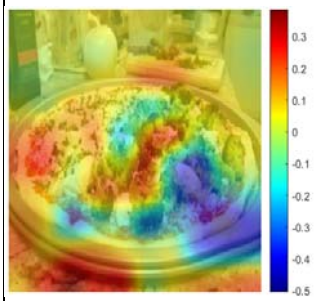
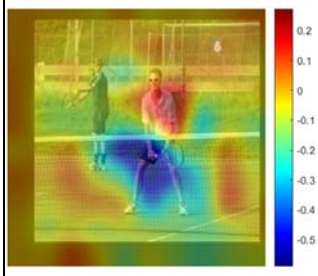


شکل (۶) نتایج پارامترهای ارزیابی شده در طول فرآیند یادگیری در مدل ارائه شده

سراسر تصویر حرکت می‌دهد تا احتمال کلاس‌های مشخص شده را محاسبه کند.

این روش می‌تواند بحرانی‌ترین مناطق تصویر را برای طبقه‌بندی برجسته کند. چند نمونه استفاده از تابع حساسیت انسداد و نواحی مهم تصویر که ویژگی‌های ضروری‌تری را برای شبکه فراهم می‌کنند ارائه شده است. استفاده از تابع حساسیت انسداد به ما کمک می‌کند تا ویژگی‌های استفاده شده توسط شبکه را بهتر درک کنیم و بینشی در مورد دلایل طبقه‌بندی نادرست تصاویر ارائه کنیم.

مقایسه توصیف‌های پیش‌بینی‌شده و توصیف‌های اصلی تصاویر در مجموعه دادگان موجود نشان می‌دهند که روش CapsNet + CNN + ELMo، یعنی بهترین مدل توصیف‌گر در این مقاله، می‌تواند توصیف‌هایی برای تصاویر ارائه دهد که تشابه بیشتر به توصیف‌های واقعی ارائه‌شده توسط انسان داشته باشد.

	توصیف واقعی: 1- 'A group of giraffes Standing up against a dirt wall in front of a crowd of children.', 2- 'A large group of people holding their arms out to feed giraffes.', 3- 'A group of children feeding two giraffes at a zoo.', 4- 'A group of people trying to feed giraffes at the zoo', 5- 'Zoo scene of children at zoo near giraffes, attempting to pet or feed them. توصیف پیش‌بینی شده: 'a giraffe is walking in the grass eeee '	
	توصیف واقعی: 1-'A man on a court swinging a tennis racket.', 2-'A tennis player is looking up to the sky.', 3-'A black and white photo of a man playing tennis.', 4- 'Black and white photograph of a man playing tennis', 5-'A man swinging a tennis racket in a t-shirt' توصیف پیش‌بینی شده: 'a man is holding a tennis racket on a tennis court eeee'	
	توصیف واقعی: 1-'a big purple bus parked in a parking spot', 2- 'A purple bus can't be missed on the city streets. ', 3-'a big purple public bus called south tyne', 4-'A city bus drives through a city area.', 5-'City bus driving through pedestrian saturated area near crosswalk.' توصیف پیش‌بینی شده: 'a double decker bus is parked on the street eeee '	
	توصیف واقعی: 1-'A pizza is prepared with cheese, tomato sauce, and broccoli. ', 2-'A pizza is split down the middle is shown.', 3-'A pizza covered with lots of broccoli sitting on a kitchen counter.', 4-'A half broccoli half cheese pizza that needs to be cooked. ', 5-'An unbaked pizza with cheese and herbs on one half and broccoli on the second'. توصیف پیش‌بینی شده: 'a pizza with a plate of food eeee'	
	توصیف واقعی: 1-'Two people standing on a tennis court playing tennis', 2- 'Two mean are playing tennis and both are wearing sunglasses. ', 3- 'Two people playing a game of tennis with rackets.', 4- 'two tennis players on a court with rackets', 5-'Two men with tennis rackets playing doubles tennis.' توصیف پیش‌بینی شده: 'a man is holding a tennis racket eeee'	

شکل (۷) : نتایج کیفی حاصل از مدل پیشنهادی

۷- نتیجه‌گیری

در این مقاله، برای توصیف محتوای تصویر یک چارچوب رمزگذار-رمزگشا ارائه شد که برای دستیابی به این هدف به پیاده‌سازی معماری‌های مختلف می‌پردازد. در بخش رمزگذار از یک شبکه کپسولی همراه با یک شبکه پیچشی به‌عنوان استخراج‌کننده اطلاعات معنایی از محتوای تصویر استفاده می‌کند. در بخش رمزگشا نیز از یک شبکه بازگشتی سه لایه استفاده می‌شود که با دریافت اطلاعات معنایی از بخش رمزگذار و بر مبنای اطلاعات متنی یادگیری‌شده در فاز یادگیری به پیش‌بینی محتوای تصویر می‌پردازد. کلیه مدل‌های ارائه شده بر روی مجموعه داده MS-COCO آموزش دیدند و بر اساس معیارهای ارزیابی BLEU (1- 4)، ROUGE و METEOR ارزیابی شدند. مقایسه نتایج مدل‌های ارائه‌شده با مدل‌های پایه نشان داد که استفاده از شبکه کپسولی همراه با شبکه پیچشی به‌عنوان استخراج‌کننده ویژگی در فاز رمزگذاری و همچنین به‌کارگیری روش ELMo برای تولید بردارهای تعبیه‌سازی کلمه‌های درون متن، در اکثر معیارهای ارزیابی منجر به نتایج قابل قبولی شد. در آینده قصد داریم استفاده از مبدل‌ها در فاز رمزگشایی و تاثیرات آنها برا تولید توصیف‌های بامعناتر برای تصاویر مورد بررسی قرار دهیم.

مراجع

[1] Y. Wei, L. Wang, H. Cao, M. Shao, and C. Wu, "Multi-attention generative adversarial network for image captioning," *Neurocomputing*, vol. 387, pp. 91–99, 2020.

[2] M. Zamiri, T. Bahraini, and H. S. Yazdi, "MVDF-RSC: Multi-view data fusion via robust spectral clustering for geo-tagged image tagging," *Expert Syst. Appl.*, vol. 173, p. 114657, 2021.

[3] R. Rad and M. Jamzad, "A multi-view-group non-negative matrix factorization approach for automatic image annotation," *Multimed. Tools Appl.*, vol. 77, no. 13, pp. 17109–17129, 2018.

[4] M. Zamiri and H. S. Yazdi, "Image annotation based on multi-view robust spectral clustering," *J. Vis. Commun. Image Represent.*, vol. 74, p. 103003, 2021.

[5] J. Asawa, M. Deshpande, S. Gaikwad, and R. Toshniwal, "Caption recommendation system," *United Int. J. Res. Technol.*, vol. 2, pp. 4–9, 2021.

[6] M. Javidi and M. Jampour, "A deep learning framework for text-independent writer identification," *Eng. Appl. Artif. Intell.*, vol. 95, p. 103912, 2020.

[7] H. Liu, G. Wang, T. Huang, P. He, M. Skitmore, and X. Luo, "Manifesting construction activity scenes via image captioning," *Autom. Constr.*, vol. 119, p. 103334, 2020.

[8] J. Wang, W. Wang, L. Wang, Z. Wang, D. D. Feng, and T. Tan, "Learning visual relationship and context-aware attention for image captioning," *Pattern Recognit.*, vol. 98, p. 107075, 2020.

[9] M. D. Z. Hossain, F. Sohel, M. F. Shiratuddin, and H. Laga, "A comprehensive survey of deep learning for image captioning," *ACM Comput. Surv.*, vol. 51, no. 6, pp. 1–36, 2019.

[10] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," *Proc. IEEE Conf. Comput. Vis. pattern Recognit.*, 2015.

[11] S. Bai and S. An, "A survey on automatic image caption generation," *Neurocomputing*, vol. 311, pp. 291–304, 2018.

[12] A. Kumar and S. Goel, "A survey of evolution of image captioning techniques," *Int. J. Hybrid Intell. Syst.*, vol. 14, no. 3, pp. 123–139, 2017.

[13] S. Sabour, N. Frosst, and G. E. Hinton, "Dynamic routing between capsules," *NIPS*, 2017.

[14] X. Ai, J. Zhuang, Y. Wang, P. Wan, and Y. Fu, "ResCaps: an improved capsule network and its application in ultrasonic image classification of thyroid papillary carcinoma," *Complex Intell. Syst.*, pp. 1–9, 2021.

[15] G. E. Hinton, S. Sabour, and N. Frosst, "Matrix capsules with EM routing," 2018.

[16] A. U. Haque, S. Ghani, and M. Saeed, "Image Captioning With Positional and Geometrical Semantics," *IEEE Access*, vol. 9, pp. 160917–160925, 2021.

[17] J. Yang and J. F. Coughlin, "In-vehicle technology for self-driving cars: Advantages and challenges for aging drivers," *Int. J. Automot. Technol.*, vol. 15, no. 2, pp. 333–340, 2014.

[18] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.

[19] Y. Jia, M. Salzmann, and T. Darrell, "Learning cross-modality similarity for multinomial data," in *Computer Vision (ICCV), 2011 IEEE International Conference on*, 2011, pp. 2407–2414.

[20] K. Barnard, P. Duygulu, D. Forsyth, N. de Freitas, D. M. Blei, and M. I. Jordan, "Matching words and pictures," *J. Mach. Learn. Res.*, vol. 3, no. Feb, pp. 1107–1135, 2003.

[21] P. Duygulu, K. Barnard, J. F. G. de Freitas, and D. A. Forsyth, "Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary," in *European conference on computer vision*, 2002, pp. 97–112.

[22] A. Farhadi *et al.*, "Every picture tells a story: Generating sentences from images," in *European conference on computer vision*, 2010, pp. 15–29.

[23] R. Socher and L. Fei-Fei, "Connecting modalities: Semi-supervised segmentation and annotation of images using unaligned text corpora," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, 2010, pp. 966–973.

architectures," in *International Conference on Machine Learning*, 2015, pp. 2342–2350.

[37] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: Lessons learned from the 2015 mscoco image captioning challenge," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 652–663, 2017.

[38] T. Ojala, M. Pietikäinen, and T. Mäenpää, "Gray scale and rotation invariant texture classification with local binary patterns," in *European Conference on Computer Vision*, 2000, pp. 404–420.

[39] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.

[40] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, 2005, vol. 1, pp. 886–893.

[41] B. E. Boser, I. M. Guyon, and V. N. Vapnik, "A training algorithm for optimal margin classifiers," in *Proceedings of the fifth annual workshop on Computational learning theory*, 1992, pp. 144–152.

[42] R. Bernardi *et al.*, "Automatic description generation from images: A survey of models, datasets, and evaluation measures," *J. Artif. Intell. Res.*, vol. 55, pp. 409–442, 2016.

[43] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 2002, pp. 311–318.

[44] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Text summarization branches out*, 2004, pp. 74–81.

[45] S. Banerjee and A. Lavie, "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments," in *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, 2005, pp. 65–72.

[46] M. E. Peters *et al.*, "Deep contextualized word representations."

[47] T.-Y. Lin *et al.*, "Microsoft coco: Common objects in context," in *European conference on computer vision*, 2014, pp. 740–755.

[48] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3128–3137.

[49] Q. Wu, C. Shen, P. Wang, A. Dick, and A. Van Den Hengel, "Image captioning and visual question answering based on attributes and external knowledge," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 6, pp. 1367–1381, 2017.

[50] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J.

[24] B. Z. Yao, X. Yang, L. Lin, M. W. Lee, and S.-C. Zhu, "I2t: Image parsing to text description," *Proc. IEEE*, vol. 98, no. 8, pp. 1485–1508, 2010.

[25] A. Gupta, P. Srinivasan, J. Shi, and L. S. Davis, "Understanding videos, constructing plots learning a visually grounded storyline model from annotated videos," in *Computer vision and pattern recognition, 2009. CVPR 2009. IEEE conference on*, 2009, pp. 2012–2019.

[26] S. Li, G. Kulkarni, T. Berg, A. Berg, and Y. Choi, "Composing simple image descriptions using web-scale n-grams," in *Proceedings of the Fifteenth Conference on Computational Natural Language Learning*, 2011, pp. 220–228.

[27] P. Kuznetsova, V. Ordonez, A. C. Berg, T. L. Berg, and Y. Choi, "Collective generation of natural image descriptions," in *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, 2012, pp. 359–368.

[28] G. Kulkarni *et al.*, "Babytalk: Understanding and generating simple image descriptions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 12, pp. 2891–2903, 2013.

[29] Y. Yang, C. L. Teo, H. Daumé III, and Y. Aloimonos, "Corpus-guided sentence generation of natural images," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2011, pp. 444–454.

[30] M. Mitchell *et al.*, "Midge: Generating image descriptions from computer vision detections," in *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, 2012, pp. 747–756.

[31] S. Lawrence, C. L. Giles, A. C. Tsoi, and A. D. Back, "Face recognition: A convolutional neural-network approach," *IEEE Trans. Neural Networks,* vol. 8, no. 1, pp. 98–113, 1997.

[32] B. Mandal, S. Ghosh, R. Sarkhel, N. Das, and M. Nasipuri, "Using dynamic routing to extract intermediate features for developing scalable capsule networks," in *2019 Second International Conference on Advanced Computational and Communication Paradigms (ICACCP)*, 2019, pp. 1–6.

[33] J. J. Hopfield, "Neural networks and physical systems with emergent collective computational abilities," *Proc. Natl. Acad. Sci.*, vol. 79, no. 8, pp. 2554–2558, 1982.

[34] K. Cho *et al.*, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," *arXiv Prepr. arXiv1406.1078*, 2014.

[35] K. Greff, R. K. Srivastava, J. Koutník, B. R. Steunebrink, and J. Schmidhuber, "LSTM: A search space odyssey," *IEEE Trans. neural networks Learn. Syst.*, vol. 28, no. 10, pp. 2222–2232, 2017.

[36] R. Jozefowicz, W. Zaremba, and I. Sutskever, "An empirical exploration of recurrent network

Dean, "Distributed representations of words and phrases and their compositionality," *Adv. Neural Inf. Process. Syst.*, vol. 26, 2013.

[51] J. Aneja, A. Deshpande, and A. G. Schwing, "Convolutional image captioning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 5561–5570.

[52] J. H. Tan, C. S. Chan, and J. H. Chuah, "Image Captioning with Sparse Recurrent Neural Network," *arXiv Prepr. arXiv1908.10797*, 2019.

[53] W. Zhang, S. Tang, J. Su, J. Xiao, and Y. Zhuang, "Tell and guess: cooperative learning for natural image caption generation with hierarchical refined attention," *Multimed. Tools Appl.*, vol. 80, no. 11, pp. 16267–16282, 2021.

[54] J. Yu, J. Li, Z. Yu, and Q. Huang, "Multimodal transformer with multi-view visual representation for image captioning," *IEEE Trans. circuits Syst. video Technol.*, vol. 30, no. 12, pp. 4467–4480, 2019.

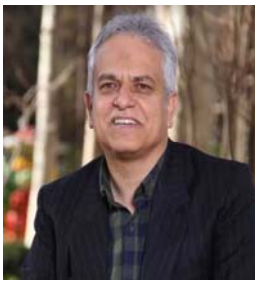
[55] J. Li, N. Xu, W. Nie, and S. Zhang, "Image Captioning with multi-level similarity-guided semantic matching," *Vis. Informatics*, vol. 5, no. 4, pp. 41–48, 2021.

[56] P. Anderson *et al.*, "Bottom-up and top-down attention for image captioning and visual question answering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 6077–6086.

[57] W. Jiang, L. Ma, Y.-G. Jiang, W. Liu, and T. Zhang, "Recurrent fusion network for image captioning," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 499–515.

[58] S. Yan, Y. Xie, F. Wu, J. S. Smith, W. Lu, and B. Zhang, "Image captioning via hierarchical attention mechanism and policy gradient optimization," *Signal Processing*, vol. 167, p. 107329, 2020.

شیما جوانمردی مدرك کارشناسی خود را در رشته مهندسی کامپیوترگرایش نرم افزار از دانشگاه جهرم در سال۱۳۹۰ و مدرك کارشناسی ارشد خود را در همان رشته در سال ۱۳۹۵ از دانشگاه یزد دریافت کرد. وي در حال حاضر دانشجوي دکتری رشته مهندسی کامپیوتر-هوش‌مصنوعی و رباتیک در دانشگاه یزد بوده و به عنوان محقق در دانشگاه لایدن هلند مشغول به فعالیت می‌باشد. زمینه‌های پژوهشی مورد علاقه وی پردازش تصویر، یادگیری ماشین و پردازش زبان طبیعی است.



علی‌محمد لطیف مدرك کارشناسی و کارشناس ارشد خود را در رشته الکترونیک به ترتیب از دانشگاه صنعتی اصفهان و دانشگاه امیرکبیر تهران در سال‌های ۱۳۷۲ و ۱۳۷۵ دریافت نمود و مدرك دکتري خود را از دانشگاه اصفهان در سال ۱۳۹۰ در رشته مهندسی کامپیوتر-هوش مصنوعی اخذ نمود. ایشان در حال حاضر به‌عنوان دانشیار دانشکده مهندسی کامپیوتر دانشگاه یزد فعالیت دارد. زمینه‌های پژوهشی مورد علاقه ایشان پردازش تصویر، مخفی نگاری داده و رمزنگاری است.

محمدتقی صادقی تحصیلات خود را در مقطع کارشناسی و کارشناسی ارشد در رشته برق-مخابرات به ترتیب از دانشگاه صنعتی شریف و دانشگاه تربیت مدرس تهران در سال‌های ۱۳۷۰ و ۱۳۷۳ دریافت نمود. نامبرده از سال ۱۳۷۴ الی ۱۳۷۸در دانشگاه یزد مشغول به کار بود. پس از آن دوره دکتری مهندسی برق-مخابرات را در دانشگاه ساری انگلستان آغاز کرده و در سال ۱۳۸۱ موفق به اخذ مدرك دکتری خود از آن دانشگاه شد. ایشان در حال حاضر به-عنوان دانشیار دانشکده مهندسی برق دانشگاه یزد فعالیت دارد. زمینه‌های پژوهشی مورد علاقه ایشان بازشناسایی آماری الگو، پردازش‌تصویر و بینایی‌ماشین است.