دانشگاه صنعتی
امیرکبیر
(پلی تکنیک تهران)

پیشنهاد پروژه تحصیلات تکمیلی
(رساله کارشناسی ارشد و دکترا)*

شماره:
تاریخ:

فرم پروژه تحصیلات تکمیلی ۱

۱- مشخصات دانشجو

نام و نام خانوادگی: احمد اسدی
شماره دانشجوئی: ۹۴۱۳۱۰۹۱
رشته تحصیلی: هوش مصنوعی و رباتیکز
دانشکده: مهندسی کامپیوتر و فن‌آوری اطلاعات
آدرس: تهران، میدان راه‌آهن، جوادیه، خیابان نوری، کوچه ندایی، پلاک ۱۹
تلفن : ۰۹۳۶۸۶۰۲۲۸۷ مقطع: کارشناسی ارشد

۲-مشخصات استاد راهنما

نام و نام خانوادگی: رضا صفابخش
سمت، مرتبه علمی و محل خدمت: استاد
آدرس: دانشگاه صنعتی امیرکبیر، دانشکده مهندسی کامپیوتر و فن‌آوری اطلاعات
تلفن: ۶۴۵۴۲۷۲۸

۳-مشخصات استاد مشاور

نام و نام خانوادگی:
سمت، مرتبه علمی:
تلفن :

۴-عنوان پایان نامه یا رساله

فارسی: تولید خودکار شرح تصاویر با استفاده از شبکه‌های عصبی کانولوشنی عمیق
انگیسی:Automatic Image Captioning using Deep Convolutional Neural Networks

نوع پروژه: کاربردی ■ بنیادی ☐ توسعه ای ☐ **تعداد واحد** ۶

۵- خلاصه پایان نامه: (مسئله فرضیات، هدف از اجراء، توجیه ضرورت انجام طرح)

با توجه به افزایش چشم‌گیر تعداد تصاویر مورد استفاده کاربران در فضاهای مجازی و همین‌طور با در نظر گرفتن گرایش روزافزون کاربران به ذخیره‌سازی تصاویر در رایانه‌های شخصی، مساله مدیریت این تصاویر و یافتن تصاویر خاص بین مجموعه تصاویر موجود، از اهمیت زیادی برخورداز است و نظر پژوهشگران زیادی را به خود جلب کرده است. گام اساسی در این راستا، دست‌یابی به سامانه‌ای است که قادر به تولید خودکار شرح برای تصاویر باشد. شرح این تصاویر، که در قالب جملات زبان طبیعی ارائه می‌شود، باید شرایط زیر را داشته باشد.

۱. توصیف صحیح کلیت[1] تصویر

شرح تولید شده توسط برنامه، باید تصویر را به صورت کلی توصیف کند. این شرح باید با تصویر مربوطه سازگار باشد؛ به این معنی که عبارات موجود در جمله باید اشیاء و افراد موجود در صحنه، ارتباط مکانی آن‌ها با یک‌دیگر و حالت هرکدام را توصیف کند.

۲. صحت جملات، به لحاظ دستوری

جملات تولید شده، هر کدام، باید مطابق با دستور زبان[2] معیار باشند.

۳. صحت جملات، به لحاظ معنایی

هرکدام از جملاتی که در شرح تولید شده به‌کار گرفته شده‌اند، باید به لحاظ معنایی، صحیح و کامل باشند.

در این پروژه قصد داریم، سامانه‌ای ارائه دهیم که قادر به تولید شرح کلی بر تصاویر باشد. دو دیدگاه اساسی در میان پژوهش‌های مشابه، به شرح ذیل، مشاهده می‌شود:

۱. یافتن نقاط توجه[3] در تصاویر و تولید جملات توصیف‌گر اشیاء مستقر در این نقاط، به طوری‌که توصیف شیء مستقر در نقطه توجه و اشیا مرتبط با آن در جملات تولیدی توصیف شده باشند.

[1] Holistic
[2] Grammar
[3] Point of Attention

۲. تولید شرح کلی بر تصاویر. در این دیدگاه، درک کلی صحنه[۴] و تولید جملات توصیفی مرتبط با آن، حائز اهمیت است.

به علاوه، شرح کوتاه تولید شده در این پروژه، به معنی تولید جملاتی است که مستقیما به توصیف صحنه، اشیاء موجود در صحنه و روابط بین آن‌ها می‌پردازند.

به طور کلی، دو چالش عمده در این پژوهش مورد توجه قرار خواهد گرفت:

۱. توصیف صحنه باید دقیق باشد؛ به این معنی که اشیاء موجود در صحنه باید کاملا از هم تفکیک و هرکدام باید به‌درستی دسته‌بندی شوند. تصویر توصیف‌شده، باید در قالب مناسبی بازنمایی شود که بتوان به راحتی از آن برای تولید جمله استفاده نمود.

۲. جملات تولیدشده برای شرح تصویر، باید به لحاظ دستور زبان، به لحاظ املایی و نیز به لحاظ معنایی صحیح بوده و همین‌طور باید با تصویر مرتبط خود سازگار باشند و آن را به درستی و با دقت شرح دهند.

۶-کلمات کلیدی فارسی: تولید خودکار شرح تصاویر، تولید جملات زبان طبیعی، شبکه‌های عصبی عمیق کانولوشنی

کلمات کلیدی انگلیسی: Automatic Image Captioning, Generating Sentences from Natural Language, Deep Convolutional Neural Networks

۷- مدت زمان اجرای پایان نامه به ماه: ۱۲ ماه **تاریخ شروع: مهرماه ۱۳۹۵**

۸- مراحل اجرای پایان نامه	۱	۲	۳	۴	۵	۶	۷	۸	۹	۱۰	۱۱	۱۲
تهیه مراجع لازم	*											
بررسی جامع سوابق موضوع	*	*	*									
فاز اول: پیش‌پردازش تصاویر				*	*							
فاز دوم: استخراج ویژگی‌ها					*							
فاز سوم: تشخیص اشیا موجود و ارتباط آن‌ها						*	*					
فاز چهارم: تناظر اشیا و عبارات مربوطه							*					
فاز پنجم: تولید جملات کامل								*	*			
فاز ششم: نتیجه‌گیری و محاسبه کارایی										*	*	
فاز هفتم: تدوین نتایج و تنظیم پایان‌نامه											*	*

۹- روش پژوهش و تکنیک های اجرائی:

فرآیند تولید خودکار شرح بر تصاویر، از دو مرحله کلی تشکیل می‌شود. تصویر ورودی ابتدا مورد پیش‌پردازش قرار می‌گیرد تا برای انجام عملیات مختلف، آماده شود. در مرحله اول فرآیند، که شامل تشخیص اشیاء، ارتباط بین اشیاء و همین‌طور مشخصات صحنه می‌شود، پس از قطعه‌بندی تصویر و تشخیص اشیاء موجود در صحنه، اقدام به دسته‌بندی اشیاء با استفاده از شبکه‌های عصبی کانولوشنی عمیق می‌شود. پس از یافتن اشیاء موجود در صحنه، یافتن ارتباط بین این اشیا و مشخصات صحنه از اهمیت زیادی برخوردار است.

در انتهای این مرحله، تصویر ورودی را می‌توان با لیستی از برچسب‌ها و ویژگی‌های کمک‌کننده به توصیف تصویر، بازنمایی نمود. بردار ویژگی تولیدشده در مرحله قبل، به عنوان ورودی به بخش تولید جملات زبان طبیعی داده می‌شود تا جملات توصیف‌کننده تصویر ورودی با توجه به این بردار ویژگی‌ها، تولید شود.

جزئیات بیشتر در مورد روش پژوهش در بخش توضیحات آورده شده است.

۱۰-سابقه علمی و فهرست منابع:

پژوهش‌های موجود در این زمینه را می‌توان به طور کلی به دو قسمت تقسیم نمود.

۱. پژوهش‌های مرتبط با درک صحنه

۲. پژوهش‌های مرتبط با تولید جملات زبان طبیعی

یکی از انگیزه‌های اولیه در پدیداری ایده تولید شرح خودکار تصاویر، با الهام از پژوهش‌های صورت گرفته در زمینه ترجمه ماشین[۵] ایجاد

[۴] Holistic Scene Understanding

شده است[۱]. پژوهش‌های موجود در زمینه ترجمه ماشین، با هدف ترجمه جملات یک زبان به زبان دیگر به طور خودکار، انجام شده‌اند. در این راستا، یک جمله از زبان مبدأ، با به‌کارگیری روش‌های مختلفی تبدیل به یک بردار ویژگی[۶] می‌شود که مشخصه‌های اصلی جمله اولیه را نمایش می‌دهد. سپس بردار ویژگی حاصل با اعمال روش‌های گوناگون دیگری، تبدیل به یک جمله از زبان مقصد می‌گردد که در آن تمام ویژگی‌های موجود در بردار ویژگی بیان شده‌اند.

با توجه به فرآیند مذکور، اگر به جای جمله زبان مبدأ، یک تصویر به بردار ویژگی، تبدیل و سپس با استفاده از روش‌های موجود قبلی، بردار ویژگی به جمله زبان مقصد ترجمه شود، معادل این است که برای تصویر ورودی، شرحی به طور خودکار تولید شده است.

شرح خودکار تصاویر، توجه پژوهشگران بسیار زیادی را به خود جلب کرده است و فعالیت‌های مرتبط متنوع و متعددی انجام شده است. به طور کلی فرآیند مورد استفاده، از دو بخش اساسی تشکیل می‌گردد:

۱. بازنمایی تصویر با استفاده از بردار ویژگی

۲. تولید جملات صحیح زبانی که بیان‌کننده بردار ویژگی تصویر باشند.

پژوهش‌های متنوعی برای بهبود نتایج هرکدام از چالش‌های بالا انجام شده است که در ادامه به بررسی برخی از آن‌ها خواهیم پرداخت.

جزئیات بیشتر در مورد ادبیات و کارهای گذشته در بخش توضیحات آورده شده است.

مراجع

[۱] Xu, K., Ba, J., Kiros, R., Courville, A., Salakhutdinov, R., Zemel, R., & Bengio, Y. (2015). Show, attend and tell: Neural image caption generation with visual attention. arXiv preprint arXiv:1502.03044.

[۲] Fidler, S., Sharma, A., & Urtasun, R. (2013). A sentence is worth a thousand pixels. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 1995-2002).

[۳] Karpathy, A., & Fei-Fei, L. (2015). Deep visual-semantic alignments for generating image descriptions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 3128-3137).

[۴] Hodosh, M., Young, P., & Hockenmaier, J. (2013). Framing image description as a ranking task: Data, models and evaluation metrics. Journal of Artificial Intelligence Research, 853-899.

[۵] Kuznetsova, P., Ordonez, V., Berg, A. C., Berg, T. L., & Choi, Y. (2012, July). Collective generation of natural image descriptions. In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1 (pp. 359-368). Association for Computational Linguistics.
Chicago

[۶] Gupta, A., & Mannem, P. (2012, November). From image annotation to image description. In Neural information processing (pp. 196-204). Springer Berlin Heidelberg.
Chicago

[۷] Sutskever, I., Martens, J., & Hinton, G. E. (2011). Generating text with recurrent neural networks. In Proceedings of the 28th International Conference on Machine Learning (ICML-11) (pp. 1017-1024).

[۸] Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002, July). BLEU: a method for automatic evaluation of machine translation. In Proceedings of the 40th annual meeting on association for computational linguistics (pp. 311-318). Association for Computational Linguistics.
Chicago

[9] Rashtchian, C., Young, P., Hodosh, M., & Hockenmaier, J. (2010, June). Collecting image annotations using Amazon's Mechanical Turk. In Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk (pp. 139-147). Association for Computational Linguistics.

[10] Young, P., Lai, A., Hodosh, M., & Hockenmaier, J. (2014). From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. Transactions of the Association for Computational Linguistics, 2, 67-78.

[11] Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, & Zitnick, C. L. (2014). Microsoft coco: Common objects in context. In Computer Vision–ECCV 2014 (pp. 740-755). Springer International Publishing.

۱۱- وسایل و تجهیزات مورد نیاز:

[۵] Machine Translation

[۶] Feature Vector

۱۲- اعتبار اجرای پایان نامه و نحوه تامین آن (ریالی و ارزی)

عنوان هزینه	ریالی	ارزی
هزینه پرسنلی وسایل و مواد مسافرت (داخل و خارج) سایر هزینه ها		
جمع کل (هزینه ها تا سقف ۲/۰۰۰/۰۰۰ ریال قابل پرداخت می باشد)		

۱۳- نظریه استاد راهنما:

امضاء

۱۴- نظریه مسئول تحصیلات تکمیلی دانشکده:

امضاء

۱۵- رئیس دانشکده :

امضاء

۱۶-تعهدنامه دانشجو:

اینجانب دانشجوی پروژه متعهد می شوم که در مدت اجرای پروژه بطور تمام وقت انجام وظیفه نموده و بدون اطلاع معاونت پژوهشی دانشگاه از مرخصی تحصیلی استفاده ننمایم و همچنین اطلاع دارم که کلیه نتایج و حقوق حاصله از این پروژه متعلق به دانشگاه بوده و مجاز نیستم بدون موافقت دانشگاه اطلاعاتی را در رابطه با پروژه به دیگری واگذار نمایم.

نام و امضاء دانشجو

۱۷-نظریه شورای تحصیلات تکمیلی دانشگاه :

امضاء

تاریخ

۱۸-**سایر توضیحات :**

روش پیشنهادی

فرآیند تولید خودکار شرح بر تصاویر، از دو مرحله کلی تشکیل می‌شود.
تشخیص اشیاء موجود در صحنه و ارتباط بین آن‌ها (درک صحنه)
تولید جملات توصیفی برای شرح تصویر

درک صحنه

تصویر ورودی ابتدا مورد پیش‌پردازش قرار می‌گیرد تا برای انجام عملیات مختلف، آماده شود. در مرحله اول فرآیند، که شامل تشخیص اشیاء، ارتباط بین اشیاء و همین‌طور مشخصات صحنه می‌شود، پس از قطعه‌بندی تصویر و تشخیص اشیاء موجود در صحنه، اقدام به دسته-بندی اشیاء با استفاده از شبکه‌های عصبی کانولوشنی عمیق می‌شود. پس از یافتن اشیاء موجود در صحنه، یافتن ارتباط بین این اشیا و مشخصات صحنه از اهمیت زیادی برخوردار است. ارتباط مکانی بین اشیا موجود با در نظر گرفتن مکان قرارگیری اشیا نسبت به هم مشخص می‌شود.

قطعه‌بندی تصاویر، فرآیند تشخیص و تمایز اشیاء مختلف را آسان‌تر می‌کند. تشخیص اشیا مختلف و دسته‌بندی آن‌ها می‌تواند با استفاده از یادگیری چند نمونه‌ای[۷] صورت می‌گیرد. همین‌طور با استخراج ویژگی‌های مختلف، می‌توان صفت‌های مختلفی مانند رنگ، اندازه و مکان هرکدام از اشیا را، تشخیص داد.

برای این منظور، با استفاده از یک شبکه عصبی عمیق کانولوشنی، بخش‌های مختلف تصویر را برچسب‌گذاری می‌نماییم. این شبکه عصبی عمیق کانولوشنی، با استفاده از جملات توصیفی موجود در مجموعه‌دادگان و تصویر مربوطه آن‌ها آموزش داده می‌شود. پس از آموزش، این شبکه قادر خواهد بود تا با دریافت یک تصویر،برچسب‌های مختلف مربوط به بخش‌های گوناگون تصویر را تشخیص دهد.

هر کدام از برچسب‌های تشخیص داده شده توسط شبکه عصبی، متناظر خواهد بود با یک عبارت توصیفی زبانی. کنار هم قرار گرفتن این عبارات به درک صحنه به نمایش گذاشته شده در تصویر منتهی خواهد شد.

در انتهای این مرحله، تصویر ورودی را می‌توان با لیستی از برچسب‌ها و ویژگی‌های کمک‌کننده به توصیف تصویر، بازنمایی نمود. بردار ویژگی تولیدشده در این مرحله، به عنوان ورودی به بخش تولید جملات توصیفی داده می‌شود تا جملات توصیف‌کننده تصویر ورودی با توجه به این بردار ویژگی‌ها، تولید شود.

تولید جملات توصیفی

در این بخش،با استفاده از یک شبکه عصبی بازگشتی، بردار ویژگی مشخص شده در مرحله قبل را تبدیل به یک یا چند جمله زبان طبیعی می‌نماییم. شبکه عصبی بازگشتی مورد استفاده در این بخش با مجموعه جملات موجود در مجموعه‌دادگان، آموزش داده می‌شود. به این طریق، این شبکه قادر خواهد بود با وجود بخش‌هایی از جمله و با کنار هم قرار دادن محتمل‌ترین کلمات، اقدام به تکمیل جمله نماید.

مروری بر ادبیات و کارهای گذشته

یکی از انگیزه‌های اولیه در پدیداری ایده تولید شرح خودکار تصاویر، با الهام از پژوهش‌های صورت گرفته در زمینه ترجمه ماشین[۸] ایجاد شده است[۱]. پژوهش‌های موجود در زمینه ترجمه ماشین، با هدف ترجمه جملات یک زبان به زبان دیگر به طور خودکار، انجام شده‌اند. در این راستا، یک جمله از زبان مبدأ، با به‌کارگیری روش‌های مختلفی تبدیل به یک بردار ویژگی[۹] می‌شود که مشخصه‌های اصلی جمله اولیه را نمایش می‌دهد. سپس بردار ویژگی حاصل با اعمال روش‌های گوناگون دیگری، تبدیل به یک جمله از زبان مقصد می‌گردد که در آن تمام ویژگی‌های موجود در بردار ویژگی بیان شده‌اند.

با توجه به فرآیند مذکور، اگر به جای جمله زبان مبدأ، یک تصویر به بردار ویژگی، تبدیل و سپس با استفاده از روش‌های موجود قبلی، بردار ویژگی به جمله زبان مقصد ترجمه شود، معادل این است که برای تصویر ورودی، شرحی به طور خودکار تولید شده است.

شرح خودکار تصاویر، توجه پژوهشگران بسیار زیادی را به خود جلب کرده است و فعالیت‌های مرتبط متنوع و متعددی انجام شده است. به طور کلی فرآیند مورد استفاده، از دو بخش اساسی تشکیل می‌گردد:

۱. بازنمایی تصویر با استفاده از بردار ویژگی
۲. تولید جملات صحیح زبانی که بیان‌کننده بردار ویژگی تصویر باشند.

پژوهش‌های متنوعی برای بهبود نتایج هرکدام از چالش‌های بالا انجام شده است که در ادامه به بررسی برخی از آن‌ها خواهیم پرداخت.

[7] Multiple Instance Learning (MIL)
[8] Machine Translation
[9] Feature Vector

بازنمایی تصاویر با استفاده از بردار ویژگی

در زمینه درک صحنه‌های جامع، روش‌های گوناگونی ارائه شده‌اند. دو روند کلی در این زمینه در بین پژوهش‌های موجود به چشم می‌خورد:

۱.استفاده از مدل‌های گرافی احتمالاتی[۱۰]

در این دسته از پژوهش‌ها، با استفاده از مدل‌های گرافی احتمالاتی، در مورد حضور یا عدم حضور اشیاء مختلف در صحنه، رابطه بین اشیاء مختلف و قطعه‌بندی[۱۱] تصویر استنتاج می‌شود. به عنوان نمونه، در مقاله[۲] یک مدل میدان تصادفی شرطی[۱۲] برای تجزیه معنایی[۱۳] تصویر ارائه شده است که با استفاده از آن می‌توان در مورد حضور یا عدم حضور اشیاء مختلف به طور توام در صحنه تصمیم‌گیری نمود.

۲.استفاده از شبکه‌های عصبی عمیق

در این دسته از پژوهش‌ها، با استفاده از شبکه‌های عصبی عمیق، پس از قطعه‌بندی تصاویر، اقدام به تفکیک اشیاء مختلف در صحنه و برچسب‌گذاری هر شیء، بسته به یادگیری انجام‌شده، می‌گردد. به عنوان نمونه در مقاله[۳] یک شبکه عصبی عمیق معرفی شده است که قادر به برچسب‌گذاری اشیاء مختلف در صحنه است. برچسب‌های مورد استفاده در این پژوهش، عبارات مختلف موجود در جملات توصیف‌گر هر تصویر در مجموعه‌دادگان می‌باشد.

تولید جملات صحیح زبانی

در این‌باره، چهار رویکرد اساسی وجود دارد:

۱.بازیابی شبیه‌ترین جمله، به طور کامل، از بین جملات موجود در مجموعه‌داده

در این دسته از پژوهش‌ها، به مساله تولید جملات به عنوان یک مساله ذخیره و بازیابی نگاه شده است؛ به این معنی که پس از نگاشت تصویر به بردار ویژگی، جمله‌ای که شبیه‌ترین جمله به بردار ویژگی است، از بین جملات موجود در مجموعه‌دادگان، به عنوان جمله خروجی انتخاب می‌شود. به عنوان مثال در مقاله[۴]، پس از نگاشت تصویر به بردار ویژگی، از بین جملات توصیف‌کننده موجود در مجموعه دادگان، جمله‌ای که نزدیکترین فاصله کسینوسی به بردار ویژگی را دارد به عنوان خروجی اعلام می‌شود. استفاده از چنین روشی در مواردی که تصویر جدیدی وارد سامانه شده است، می‌تواند منجر به تولید جملاتی شود که به خوبی روابط بین اشیاء تصویر را نمایش ندهند یا جملاتی که علاوه بر نمایش روابط اشیاء موجود، روابط دیگری را نیز توصیف نمایند که مطلوب نیست.

۲.شکستن جملات موجود در مجموعه‌دادگان و استفاده از عبارات موجود در آن‌ها برای تولید جمله

برای حل مشکل روش قبل، به جای استفاده از کل جملات موجود به عنوان مجموعه جملات خروجی، می‌توان جملات موجود را به عبارات معنادار آن‌ها تجزیه نمود و از بین عبارات موجود، عباراتی را که بیان‌کننده بردار ویژگی حاصل می‌باشند به عنوان خروجی نمایش داد. در مقاله[۵] یک نمونه از کاربرد این روش مشاهده می‌شود. در این مقاله پس از بازیابی عبارات مرتبط موجود، و پس از اعمال قیود زبانی و امتیازدهی به عبارات، عباراتی را که امتیاز بدست آمده را بهینه می‌کنند به عنوان عبارات مناسب انتخاب نموده و با توجه به نقش آن‌ها در جمله، اقدام به تولید جملات جدید می‌نمایند.

۳.استفاده از کلیشه[۱۴]های ثابت

در مقاله[۶] ابتدا با تشخیص و دسته‌بندی اشیاء مختلف موجود در صحنه، اقدام به ساخت عبارات زبانی شامل نام اشیاء و صفات مربوط به آن‌ها شامل رنگ و اندازه کرده و سپس با استفاده از این عبارات و کلیشه‌های ثابت موجود، با پرکردن جاهای خالی کلیشه‌ها توسط عبارات، جملات توصیفی برای تصویر تولید می‌شوند.

۴.استفاده از شبکه‌های عصبی بازگشتی برای تولید جملات جدید

برخلاف روش‌های قبلی، که تولید جملات جدید در آن‌ها یکی از مشکلات اساسی بود، دسته جدیدی از پژوهش‌های مرتبط در زمینه تولید شرح بر تصاویر وجود دارد که برای تولید جملات زبانی از شبکه‌های عصبی مصنوعی بازگشتی[۱۵] استفاده می‌نمایند.

---

[۱۰] Probabilistic Graphical Model
[۱۱] Segmentation
[۱۲] Conditional Random Field (CRF)
[۱۳] Semantic Parsing
[۱۴] Template

به عنوان نمونه در مقاله[۳]، پس از نگاشت تصویر به بردار ویژگی، یک شبکه عصبی بازگشتی ارائه شده است که در هر گام، احتمال رخداد کلمه بعدی در جمله را با توجه به جملات قبلی تولید شده، محاسبه کرده و کلمه با بیشترین احتمال را مشخص می‌نماید. همین‌طور در مقاله[۷]، یک شبکه عصبی بازگشتی ارائه شده است که با استفاده از بهینه‌سازی بدون هسین[۱۶] آموزش می‌بیند و به منظور تولید حرف به حرف[۱۷] جمله جدید مورد استفاده قرار می‌گیرد.

مجموعه‌داده

به طور کلی، مجموعه‌داده‌های بسیار زیادی برای استفاده در زمینه تولید خودکار شرح تصاویر ارائه شده است و در دسترس می‌باشند. از این میان سه مجموعه‌داده فلیکر۸[۱۸][۹]، فلیکر۳۰[۱۹][۱۰] و اماس‌کوکو[۲۰][۱۱] از جمله مهم‌ترین مجموعه‌داده‌هایی هستند که در پژوهش-های زیادی مورد استفاده قرار گرفته‌اند. به دلیل وجود امکان مقایسه نتایج روش ارائه شده در این پروژه با پژوهش‌های مشابه، در پروژه حاضر نیز از همین سه مجموعه‌داده استفاده خواهد شد.

آزمون و ارزیابی

بخش‌های مختلفی از فرآیند را که نیاز به ارزیابی دارند، می‌توان به سه دسته کلی تقسیم‌بندی نمود:

خروجی بخش‌های مربوط به پردازش تصویر و بینایی ماشین، شامل مراحلی مانند قطعه‌بندی تصاویر، تشخیص و دسته‌بندی اشیاء که برای ارزیابی آن‌ها می‌توان از معیارهای موجود و معمول استفاده نمود.

ارزیابی میزان مناسب بودن شرح تولید شده برای تصویر که در ادامه توضیح داده خواهد شد.

مهم‌ترین و اساسی‌ترین معیار در ارزیابی شرح تولید شده، داشتن معیاری برای سنجش میزان مناسب‌بودن خروجی می‌باشد. به این منظور از یکی از معیارهایی که در زمینه ترجمه ماشینی در مقاله[۸] ارائه شده است، استفاده می‌نماییم. در این مقاله، معیاری تحت عنوان BLEU[۲۱] مطرح شده است که برای مقایسه میزان مناسب بودن دو ترجمه متفاوت از یک جمله مشترک در زبان مبدأ، به‌کار می‌رود.

برای استفاده از این معیار در ارزیابی خروجی، اگر تصویر ورودی را معادل با جمله زبان مبدأ بدانیم، به ازای تصاویر موجود، یک ترجمه در دسترس، همان شرحی است که به عنوان شرح تصویر در مجموعه‌داده موجود است و ترجمه دیگر آن، شرح تولیدی برنامه می‌باشد. با استفاده از معیار BLEU می‌توان میزان بهتر بودن ترجمه‌های تولیدی برنامه (شرح تولید شده به طور خودکار) را نسبت به ترجمه‌های موجود، شرح نوشته شده توسط عوامل انسانی که در مجموعه‌داده موجود است، سنجید.

معیار دیگر که می‌تواند به عنوان معیار خوبی برای سنجش میزان کارا بودن شرح تولید شده در نظر گرفت، میزان ترجیح شرح تولید شده به شرح موجود می‌باشد. برای استفاده از این معیار، تصویر ورودی را به همراه شرح موجود در مجموعه‌داده و شرح تولید شده سامانه به یک فرد خبره نمایش می‌دهیم. درصد شرح‌های تولیدی که توسط فرد خبره به شرح موجود ترجیح داده شده‌اند به عنوان معیاری برای میزان خوب بودن شرح تولید شده، مورد استفاده قرار می‌گیرد.

---

[۱۵] Recursive Artificial Neural Network (RNN)
[۱۶] Hessian-Free Optimization
[۱۷] Character by Character
[۱۸] Flickr8k
[۱۹] Flickr30k
[۲۰] MSCOCO
[۲۱] Bidirectional Language Evaluation Understudy