An Analysis on Recent Approaches for Image Captioning

Qazi Anwar[1], Ch V S Satyamurty[2]
[1]PG Scholar, CVR College of Engineering/IT Department, Hyderabad, India
Email: 20B81DB002@cvr.ac.in
[2]Associate Professor, CVR College of Engineering/IT Department, Hyderabad, India
Email: satatyamurty@cvr.ac.in

Abstract: **Image captioning is an interdisciplinary area that uses techniques from computer vision and natural language processing to provide a textual description of a picture. The Image captioning task is the process of understanding the scene present in the image by identifying objects and associated actions present to create a meaningful human-like caption which can be used for wide range of applications, including image retrieval, video indexing, assistive technology for the visually impaired, content-based image search, biomedicine, and autonomous cars. Formerly, Machine Learning was utilized for this purpose which will be extensive use of hand-crafted features such as Scale-Invariant Feature Transform (SIFT), Local Binary Patterns (LBP), the Histogram of Oriented Gradients (HOG), and combinations of these features. Extracting handmade characteristics from huge datasets is not straightforward or viable. Many deep learning-based techniques were later proposed. Deep Learning retrieval and template-based approaches were presented; however, both had drawbacks such as losing crucial objects. Recent breakthroughs in deep learning and natural language processing have resulted in considerable increases in image captioning system performance which involves adopting attention mechanisms, transformer-based architectures, multi modal connections, Object-Detection based encoder-decoder and many others. In this survey will explore some of the most recent techniques for image captioning, the datasets and evaluation measures that have been employed in deep learning-based automatic image captioning. The ultimate intention of this study is to act as a guide for researchers by emphasizing future directions for research work.**

Index Terms: **image captioning, computer vision, deep learning, Textual description, natural language processing.**

I. Introduction

In recent years, computer vision in image processing has achieved major advances, such as Image categorization [1] and object identification [2]. As a result, it is now feasible to automatically produce one or more phrases to comprehend the visual information of an image, a challenge known as Image Captioning. To develop natural language descriptions of images in image captioning in a way that a person can understand. Because the

images will be translate into words, image captioning is further challenging work than other computer vision tasks. Deep learning algorithms automatically learn characteristics from training data and can handle a large and diverse range of photographs and videos.

Deep learning systems used CNN for image processing. CNN harvests feature without the need for human involvement by automatically training and updating network parameters, while a classifier such as Softmax is used for classification. The previous image description approach employed CNN and RNN such as LSTM as encoder and decoder, respectively. The automatic generation of entire natural picture descriptions has a wide range of possible applications, including titles linked to news photographs, and information related from medical images, text-based image retrieval, information accessed by blind users, and human-robot interaction. After training a CNN on ImageNet [3] Vinyals et al. [4] used an LSTM kind of RNN to decode and output the caption. However, it is recommended to use a range of characteristics from different portions of the image since this strategy misses key regions of the image. The typical LSTM approach focuses on the relatively close vocabulary while ignoring the distant one. Recent improvements in image captioning paved the way for implementing attention methods, transformer-based architectures, multimodal connections, and object-detection based encoder-decoders, among other things may be credited in significant part to vision-language pre-training (the current dominant training paradigm for vision-language (VL) research.

II. Recent Approaches

Ren et al. introduced reinforcement learning for photo captioning in 2017 [5]. This technique's architecture comprises of two networks that work in tandem at each time stamp to find the most suitable phrases. The "policy network" gives regional direction and word projection on the current scenario. The "value network" provides global direction; it considers the reward offered by the reinforcement strategy and analyses the reward value for all prospective improvements to the existing situation. It uses an Actor-Critical Reinforcement learning technique to train this entire system [6].

In [7], researchers employed the YOLO object identification algorithm as the encoder component and the LSTM as the language decoder part to caption for the MSVD dataset, identify relevant frames from the clip that may be used to train the model KATNA, and delete redundant frames.

Following the extraction of relevant frames YOLO is utilized to detect objects, and the found items are recorded in a text file along with their classes and confidences. Then, by scanning three words at a time, LSTM creates a phrase demonstrating the relationship between the discovered items.

In [8], they suggested the Reflective Decoding Network (RDN) for picture captioning, which improves the long sequential modeling capability of classic caption decoders.

Received on 15-11-2023, Revised on 15-12-2023, Accepted on 14-01-2024.

The RDN centers on the target interpreting side and implementation consideration instruments in both the visual and literary spaces, in difference to prior approaches that upgraded captioning execution by improving the visual consideration instrument or by upgrading the encoder to supply a more significant middle of the road representation for the decoder. Upon accepting an input picture, the show begins with using Speedier R-CNN as an encoder to get visual information from the picture. The visual elements are subsequently sent to RDN for caption generation. It is made up of three parts: (1) Attention-based Repetitive Module, which goes to the Encoder's visual highlights (2) Reflective Consideration Module, which provides printed consideration to show the compatibility between current and past interpreting covered up states, permitting it to capture more authentic and comprehensive data for word choice (3) The Reflective Position Module, which gives relative position data for each word within the delivered caption and helps the show in seeing sentence syntactic structure. RDN can beat the long-term certainty issue in caption interpreting.

Transformer-based systems are cutting-edge in sequence modeling applications such as machine translation and language interpretation. However, its relevance to multi-modal situations such as picture captioning remains mainly unexplored. In [9] introduce a Meshed Transformer with Memory for Image Captioning. The architecture improves the language generation stage as well as the image encoding stage: the architecture learns a multidimensional representation of the relationship between the image regions by embedding learned a posteriori, and in the decoding stage it leverages mesh-like connections to leverage low-level as well as high-level functions. The two key features of this architecture (i) picture portions and their connections are encoded on several levels, with low-level and high-level relationships considered. Using permanent memory vectors, the model can learn and encode priori information when modeling these interactions. (ii) Using a multi-layer architecture, the sentence creation harnesses both low- and high-level visual associations rather than relying on a single input from the visual sensory system. This is undertaken using recognized gating mechanism that weights multi-level contributions at each stage. This results in a mesh connection schema between the encoder and decoder layers.

CLIP is employed in [10] to extract visual data from photos, and a mapping network is then used to generate a large number of context tokens. The mapping network employed is a Multi-Layer Perceptron or Transformer, which translates CLIP embeddings to language model space. To generate the image description, the language model will be trained using these context tokens. They choose a language model (GPT-2) [11] to create the next word of a caption from a collection of initial context tokens in less time than heavier architectures, and just the image encoder of CLIP (Contrastive Language-Image Pre-training) is used, with text encoder ignored.

For vision-language issues, large-scale pre-training approaches for learning cross-modal representations on image-text pairings are becoming unmistakable. In [12] Oscar which uses question labels recognized in pictures as grapple focuses to essentially ease arrangement learning was presented. The revelation that the conspicuous things in a picture may be dependably recognized and are regularly tended to within the going with content empowered our procedure. They train an Oscar demonstrate employing a open corpus of 6.5 million text-image pairings some time recently fine-tuning it on downstream assignments, coming about in unused state-of-the-arts on six well-established vision-language understanding and era errands.

[13] This research includes a deep investigation into enhancing visual representations for vision language (VL) problems, as well as the development of an enhanced object identification model to give image-centric representations. Compared to the most popular top-down and bottom-up models, the new model is larger, better suited for VL processes, and has been trained on significantly larger pre-trained corpora that contain several publicly available annotated sets of item identification data. This suggests that a far wider range of pictures and ideas may be represented by the new approach when it comes to building representations. The approach show that visual components are significant in VL processes, in contrast to prior VL research that neglected the original object detection model in favor of improving the vision and language fusion model. The new object identification model's visual characteristics were entered into OSCAR, a transformer based VL fusion model, throughout the tests, and the model was pre-trained and refined on a variety of downstream VL model tasks using an upgraded technique, OSCAR+.

The test results have seen a considerable improvement in the performance of the picture captioning challenge based on vision language pre-training (VLP) in recent years. However, most of the previous research has primarily focused on pre-training transformers of moderate sizes such as 12 or 24 layers on around 4 million photos.

In [14] LEMON, LargE-scale iMage captiONer, driven the primary observational analysis on the scaling behavior of VLP for picture captioning in this work. They utilize the cutting-edge Vinyl demonstration as a reference, which encompasses a picture highlight extractor and a transformer demonstration, and to scale the transformer up and down, with demonstration sizes extending from 13 to 675 million parameter.

[15] The unused show is greater, and superior outlined for VL assignments. It is additionally pre-trained on much bigger preparing corpora containing different straightforwardly explained Object-Identification datasets than most broadly utilized bottom-up or top-down models, permitting it to build representations of a huge set of visual objects/concepts. Though past VL inquire about centered on this work presented mPLUG, the modern vision-language establishment worldview for cross-modal comprehension /generation assignments. The long visual grouping in cross modular arrangement leads to moo computational effectiveness & data lopsidedness in most pre prepared models. To address this, mPLUG offers a progressive cross modular skip-connecting engineering that produces inter-layer alternate routes which bypass the set number of layers, permitting for time devouring full self-attention on vision side. mPLUG is pre-trained conclusion to conclusion on huge scale image-text sets with segregating and generative destinations. It produces cutting edge comes about on a wide extend of downstream VL

errands, such as picture captioning / image-text recovery / visual establishing / visual address replying. When connected to distinctive video-language assignments, mPLUG moreover shows zero-shot transferability.

[16] They suggest employing CLIP, a multimodal encoder trained on massive image-text pairings from the web, to calculate multimodal similarity and utilize it as a reward function to generate more informative and differentiated captions. They too recommend a basic fine-tuning strategy for the CLIP content encoder that does not require extra content observes to make advances in linguistic use. To survey graphic captions comprehensively, they moreover presented Fine Cap Eval, a novel dataset for caption assessment with fine-grained criteria. The proposed CLIP guided show produces more interesting captions than the CIDEr-optimized show. They too appear that the unsupervised linguistic use fine-tuning of the CLIP content encoder lightens the naive CLIP reward's degeneration issue. At last, they give human examination in which annotators exceedingly favor the CLIP remunerate over the CIDEr and MLE targets based on various parameters.

III. DataSets

A. Mscoco (Microsoft Common Objects in Context)

It is a significant dataset for segmentation, captioning, and object recognition [17]. 328K images make up the collection. The initial release of this dataset took place in 2014. It comprises three sets of 164K photos: 83K for training, 41K for validation, and 41K for testing. An extra test set of 81K photos was made available in 2015; this set contained 40K new photos in addition to all previous test images. In the year 2017, the split between training and test was changed from 83K/41K to 118K/5K. The updated split makes use of the same images and annotations. The 41K pictures in the 2015 test set are not included in the 2017 test set. An additional new, unannotated dataset of 123K images is included in the 2017 edition.

B. Flickr

A benchmark [18] collection of 8,000 photos comprised with five distinct captions that convey concise explanations of the main things and happenings. The photos were taken from six diverse Flickr bunches and don't as a rule incorporate well-known people or places but were hand-picked to demonstrate a differing quality of scenarios and circumstances. A follow-up to the prior Flickr 8k Dataset, a dataset Flickr 30k was introduced. It is a picture caption corpus comprising 158,915 crowd-sourced captions depicting 31,783 photographs depicting people locked in in standard exercises and occasions.

C. Conceptual Captions

Google's Conceptual collection [19] contains over three million photos accompanied by natural language narratives. The Conceptual Captions images and raw explanations are taken from the web and so reflect a broader range of styles. The raw descriptions are derived from the Alt-text HTML element of online photographs. The splits comprise of 3.3 million training images, 15k validation images and 12k for testing purposes.

D. Coco

COCO Captions [20] offers approximately 0.5 million captions that describe over 330,000 photos. Each picture within the preparing and approval sets will incorporate five distinct human-generated captions. The entire number of captions collected is 413,915 for 82,783 photographs in the organization of 202,520 for 40,504 images in approval, and 379,249 for 40,775 pictures in testing. There's an additional caption for each testing picture to compute human execution evaluations for comparing machine created caption scores. The whole number of captions accumulated is 1,026,459.

D. Hateful Memes

The Hateful Memes collection [21] includes over 10,000 new multimodal instances generated by Facebook AI to assist researchers in developing new algorithms to detect multimodal hate speech. This information includes several modalities, such as text and graphics, making machine comprehension challenging. The photos were licensed from Getty photos so that scholars may utilize the dataset in their studies.

E. No Caps

166,100 human-generated captions portraying 15,100 photographs from the Open photographs training and test sets contain this benchmark [22]. COCO image-caption pairings, as well as Open Pictures image-level names and question bounding boxes, make up the related preparing information. Since Public photographs has distant more classes than COCO, over 400 thing classes found in test photographs have no or few preparing captions.

F. Viz Wiz

It is a visual question answering (VQA) dataset [23] which evolved from a natural visual question answering scenario in which unsighted persons individually took a picture and recorded a spoken inquiry about it, along with ten crowd sourced replies per visual question. For this dataset there are two tasks: (1) imagining the reply to a visual address and (2) determining in case a visual address cannot be replied. There are 20,523 training image/question pairings, 205,230 pairs of training answer/answer confidence, 4,319 image/question validation pairings, 43,190 validation confidence answer/answer pairings, 8,000 image/question pairings for testing.

G. Rsicd

The Remote Sensing Image Captioning Dataset [24] is a dataset used to caption remote sensing images. The dataset includes almost 10 thousand remote sensing photos gathered from Baidu Map, Google Earth, Tianditu. MapABC, and the photos are restrained to 224X224 pixels in a diversity of dimensions. There are 10921 remote sensing photos in all, with five phrase explanations for each image.

H. Image Captioning For Visually Impaired People

This data [25] collection comprises 1600 distinct images organized into 21 major categories. The images remained chosen to represent the various scenarios or obstacles that visually impaired people will encounter in a real-world setting to assist them in a variety of ways, such as crosswalks, construction activity to sign boards, food outlets, stairs, elevators, bus terminals, wet surfaces, push buttons, money recognition, and so on.

IV. Evaluation Metrics

A. BLEU

To evaluate the quality of interpreted expressions in machine interpretation, the Bilingual Evaluation Understudy (BLEU) approach [26] is utilized. It investigates person interpretation fragment to a pool of reference interpretations of incredible interpretation quality, gauges each section score, and after that assesses the full interpretation quality. Within the domain of picture portrayal, BLEU uses a coordinating run the show as a similitude measuring approach. Utilizing the co-occurrence frequency of N-gram in both the label and the predicted caption, one may evaluate the BLEU assessment metric. To be computed are four bleu scores. The depiction express and name are separated into words by BLEU-1, which checks the occurrences of each word within the depiction sentence within the name one at a time and logs the smallest number of times a tuple shows up within the depiction sentence and name. To avoid the bias issue of the resulting description sentence being too short, compute the ratio using the description sentence and multiply the result by a penalty factor. For statistical and computational purposes, BLEU-2 separates the descriptive phrase and label into two-word 2-tuples. Typically, a maximum of four tuples are computed.

B. ROUGE

The ROUGE (recall-oriented understudy for gisting evaluation) approach [27] investigations abstracts based on the co-occurrence data of the N-tuples within the assessment abstracts. It is an evaluation method used to gauge the machine's translation fluency that is based on the recall rate of N-tuples. In order to verify the lengthiest common subsequence between the label and the captionl, ROUGE uses dynamic programming in the evaluation process. Based on the computed common subsequence, they next compute their recall to determine how similar the caption and label are. Like BLEU, the higher the ROUGE indicator's value, the higher the quality; however, it does not account for semantic depth of description or grammatical precision., ROUGE is able to capture the phrase's structure.

C. METEOR

METEOR (Metric for Evaluation of Translation with Explicit Ordering) [28] is another machine translation evaluation index. The METEOR before performing a harmonic average for a query image caption estimates recall and precision. The longer the ceaseless length of the longest common subsequence coordinated, we normally accept, the way better. Be that as it may, since the assessment metric assesses a single word, a punishment figure is utilized, where the calculate shows the number of chunks, which means the bordering requested square. METEOR fathoms the issue that BLEU does not dependably coordinate words and does not survey review. METEOR measures exact word-to-word coordinating but BLEU does in an indirect way.

D. CIDER

Consensus-based Image Description Evaluation [29] treats each phrase as an archive and calculates the cosine point of the word frequency-inverse record recurrence vector to decide the closeness between the depiction sentence and the name. At last, the result is calculated by averaging the likeness of tuples of changing lengths. Since tuples that show up more frequently within the corpus by and large carry less data, this method permits different tuples to have shifting weights. As a result, CIDER may assess graphic expressions for rightness etymologically and really.

E.SPICE

Anderson et al. [30] developed Semantic Propositional Image Caption Evaluation (SPICE) to employ graph-based semantic illustration to encode the objects, properties, and connections in the description sentence and to assess the description sentence at the semantic level. SPICE uses a dependency parser to convert the candidate and reference captions into syntactic dependencies trees. The era of the dependency tree, a rule-based method is utilized to interpret the reliance tree into a scene chart. The syntactic Reliance Tree is built particularly by three post-processing stages that streamline quantitative modifiers, investigate pronouns, and handle different things. The tree structure is at that point prepared utilizing nine fundamental dialect rules to recover the scene graph's things, connections, and properties. While Flavor can way better evaluate semantic data, it overlooks linguistic confinements and thus cannot judge sentence stream.

In Table-1 summarizes the most pertinent survey techniques.and their primary features regarding visual encoding, language modelling, and training methodologies along with their performance on the COCO Karpathy test set in terms of BLEU-4, CIDEr,, and METEOR. The strategies are sorted within the table concurring to the evaluations they learned. Strategies that take advantage of pre-training in dialect and visual perception are enhanced and acclaimed by others. In a matter of a long time, picture captioning models have accomplished exceptional results, 25.1 from a normal BLEU-4 of for the strategies utilizing worldwide CNN highlights to the normal BLEU-4 of 35.3 and 40.0 for consideration and self-attention instruments, with the latter peaking at 42.6 within the case of pre-training of vision-and-language. The way better the execution, when measured in terms of the CIDEr score, is when total and organized data approximately semantic visual concepts and their connections is included. With respect to the dialect show, the execution of LSTM-based methods with vigorous visual encoders is individually competitive with that of afterward completely mindful methods.

TABLE 1.
SUMMARY OF DL BASED IMAGE CAPTIONING MODELS

Visual Encoding					Language Model				Training Strategies				Main Results		
Model	Global	Regions	Grid	Graph	Self-Attention	RNN/ LSTM	Transformer	BERT	MLM	XE	Reinforce	VL Pre training	BLEU-4	METEOR	CIDEr
Unified VLP		✓			✓			✓		✓	✓	✓	39.5	29.3	129.3
RDN		✓				✓				✓			36.8	27.2	115.3
M^2 Transformer		✓			✓		✓			✓	✓		39.1	29.2	131.2
Universal Cap			✓		✓		✓			✓	✓	✓	40.8	30.4	143.4
CPTR					✓		✓			✓	✓		40	29.1	129.4
Oscar		✓			✓			✓	✓		✓	✓	41.7	30.6	140
LEMON		✓			✓				✓		✓	✓	42.6	31.4	145.5
Embedding Reward	✓					✓				✓	✓		30.4	25.1	93.7

V. CONCLUSIONS

The study examined the model frameworks proposed in recent years in computer vision for image captioning tasks, specifically the object detection-based approach, Reinforcement Learning, multi-modal connections, and the Vision Language Pre-training approaches and also reviewed different datasets and evaluation metrics associated with the image captioning task. Even though the interpretation of the reinforcement learning, attention methods, and object detection-based techniques are good, it is concluded that employing Vision Language Pre-training approaches for downstream tasks such as picture captioning can outperform conventional designs.

REFERENCES

[1] Philip Kinghorn, Li Zang, “a region based image caption generator with refined descriptions” , Elsiver B V, 6 july 2017, Ling Shao University Northumbria New castle NE1,United Kingdom.

[2] Priyanka Raut, Rushali A Deshmukh, “An Advanced Image Captioning using combination of CNN and LSTM”, Turkish Journal of Computer and Mathematics Education, 05 April 2021, Savitribai Phule Pune Univresity, faculty, Maharhatra/India.

[3] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “Imagenet: A large-scale hierarchical image database,” in Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 248–255, 2009.

[4] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, “Show and tell: A neural image caption generator,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2015.

[5] Zhou Ren, Xiaoyu Wang, Ning Zhang, Xutao Lv, and Li-Jia Li. 2017. Deep Reinforcement Learning-based Image Captioning with Embedding Reward. In Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR). 1151–1159

[6] L. Zhang, F. Sung, F. Liu, T. Xiang, S. Gong, Y. Yang, and T. M. Hospedales, “Actor-Critic Sequence Training for Image Captioning,” in NeurIPS, 2017

[7] Hanan Nasser Alkalouti, Dr. Mayada Ahmed, “Encoder-Decoder Model for Automatic Video Captioning Using Yolo Algorithm”, IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS), 21-24 April 2021.

[8] Ke, L., Pei, W., Li, R., Shen, X., & Tai, Y. W. (2019, October). Reflective Decoding Network for Image Captioning. 2019 IEEE/CVF International Conference on Computer Vision. https://doi.org/10.1109/iccv.2019.00898

[9] Cornia, M., Stefanini, M., Baraldi, L., & Cucchiara, R. (2019). M2: Meshed-Memory Transformer for Image Captioning. ArXiv, abs/1912.08226

[10] R. Mokady, A. Hertz, and A. H. Bermano, “Clipcap: CLIP prefix for image captioning,” Computing research repository, vol. abs/2111.09734, 2021.

[11] Alec Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, “Language models are unsupervised multitask learners,” OpenAI blog, vol. 1, 2019.

[12] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, & Jianfeng Gao. (2020). Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks.

[13] Zhang, Pengchuan & Li, Xiujun & Hu, Xiaowei & Yang, Jianwei & Zhang, Lei & Wang, Lijuan & Yejin, Choi & Gao, Jianfeng(2021).VinVL: Revisiting Visual Representations in Vision-Language Models. 5575-5584.

[14] Hu, X., Gan, Z., Wang, J., Yang, Z., Liu, Z., Lu, Y., & Wang, L. (2022, June). Scaling Up Vision-Language Pretraining for Image Captioning. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). https://doi.org/10.1109/cvpr52688.2022.01745

[15] Chenliang Li, Haiyang Xu, Junfeng Tian, Wei Wang, Ming Yan, Bin Bi, Jiabo Ye, He Chen, Guohai Xu, Zheng Cao, Ji Zhang, Songfang Huang, Fei Huang, Jingren Zhou, and Luo Si. 2022. mPLUG: Effective and Efficient Vision-Language Learning by Cross-modal Skip-connections. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 7241–7259, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

[16] Cho, J., Yoon, S., Kale, A., Dernoncourt, F., Bui, T., & Bansal, M. (2022). Fine-grained Image Captioning with CLIP Reward. Findings of the Association for Computational Linguistics: NAACL 2022.

[17] Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., & Zitnick, C. L. (2014). Microsoft COCO: Common Objects in Context. Computer Vision – ECCV 2014, 740–755. https://doi.org/10.1007/978-3-319-10602-1_48

[18] https://www.kaggle.com/datasets/hsankesara/flickr-image-dataset

[19] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual Captions: A Cleaned, Hypernymed, Image Alt-text Dataset For Automatic Image Captioning. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 2556–2565, Melbourne, Australia. Association for Computational Linguistics.

[20] Chen, Xinlei & Fang, Hao & Lin, Tsung-Yi & Vedantam, Ramakrishna & Gupta, Saurabh & Dollar, Piotr & Zitnick, C.. (2015). Microsoft COCO Captions: Data Collection and Evaluation Server.

[21] Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2020. The hateful memes challenge: detecting hate speech in multimodal memes. In Proceedings of the 34th International Conference on Neural Information Processing Systems (NIPS'20). Curran Associates Inc., Red Hook, NY, USA, Article 220, 2611–2624.

[22] Agrawal, Harsh & Desai, Karan & Chen, Xinlei & Jain, Rishabh & Batra, Dhruv & Parikh, Devi & Lee, Stefan & Anderson, Peter. (2018). Nocaps: novel object captioning at scale.

[23] Gurari, D., Li, Q., Stangl, A. J., Guo, A., Lin, C., Grauman, K., Luo, J., & Bigham, J. P. (2018, June). VizWiz Grand Challenge: Answering Visual Questions from Blind People. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. https://doi.org/10.1109/cvpr.2018.00380

[24] Lu, Xiaoqiang & Wang, Binqiang & Zheng, Xiangtao & Liu, Wei. (2017). Exploring Models and Data for Remote Sensing Image Caption Generation. IEEE Transactions on Geoscience and Remote Sensing.

[25] https://www.kaggle.com/datasets/aishrules25/automatic-image-captioning-for-visually-impaired.

[26] Papineni, K., Roukos S., Ward T., Zhu W.J.: Bleu: A method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting of the Association for Computational Linguistics, pp. 311– 318 (2002)

[27] C.-Y. Lin, “Rouge: A package for automatic evaluation of summaries,” in ACL Workshop, 2004.

[28] M. Denkowski and A. Lavie, “Meteor universal: Language specific translation evaluation for any target language,” in EACL Workshop on Statistical Machine Translation, 2014.

[29] R. Vedantam, C. L. Zitnick, and D. Parikh, “Cider: Consensus-based image description evaluation,” arXiv preprint arXiv:1411.5726, 2014.

[30] Anderson, P., Fernando B., Johnson M., Gould S.: Spice: Semantic propositional image caption evaluation. In: Proceedings of the European Conference on Computer Vision, pp. 382– 398 (2016).