Proceedings of the 3rd International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/5/20230511

To describe the content of image: The view from image captioning

Xiaohan Hou

Faculty of Math, Mathematics Studies, University of Waterloo, Waterloo, N2L 3G1

x39hou@uwaterloo.ca

Abstract. The aim of developing the technology of "image captioning," which integrates natural language and computer processing, is to automatically give descriptions for photographs by the machine itself. The work can be separated into two parts, which depends on correctly comprehending both language and images from a semantic and syntactic perspective. In light of the growing body of information on the subject, it is getting harder to stay abreast of the most recent advancements in the area of image captioning. Nevertheless, the review papers that are now available don't go into enough detail about those findings. The approaches, benchmarks, datasets, and assessment metrics currently in use for picture captioning are reviewed in this work. The majority of the field's ongoing study is concentrated on robust learning-based techniques, where deep reinforcement, adversarial learning, and attention processes all seem to be at the heart of this research area. Image captioning entails a brand-new field in research on computer vision. Generating a comprehensive natural language description for the source images is the fundamental issue of image captioning. This essay explores and evaluates earlier work on image captioning. Image captioning's application and task situations are introduced. The merits and disadvantages of each approach are explored after the analysis of the image captioning algorithms based on encoder-decoder and template structure. The assessment and baseline dataset for picture captioning are therefore shown. Ultimately, prospects for image captioning's progress are presented.

Keywords: Image captioning, semantic perspective, syntactic perspective, computer processing.

1. Introduction

People learn about the world in a variety of ways, one of which is through the analysis and acquisition of visual information. The scene images can become their own thinking. Individuals can increasingly understand their surroundings through constant accumulation. It is noteworthy that mage captioning entails a study area that turns visual information into cognitive knowledge. Its fundamental paradigm requires the functionality of two different pieces. The first step is to collect the picture's features, primarily the object's data and its location; the second is to evaluate the semantics of the picture evaluation and integrate it with the image elements to produce the image description. There is a full cognitive framework in the human mind. The brain analyses the image and examines its information as soon as it is received. It is frequently important to give computers the capability of cognitive image when they achieve image captioning. Consequently, this functionality cannot be achieved via the conventional program. Additionally, the program and the rationale to be considered are both too big,

DOI: 10.54254/2755-2721/5/20230511

and also, the conventional program is just too stiff to provide the desired result. With artificial intelligence as a foundation, the neural net method is used to bring computers closer to human cognition and raise them to child language descriptive levels. The utilization of attention processes is one method that is essential in today's modern picture captioning. Since the invention of transformers, several diverse activities, including machine translation and language modelling, have been improved. This paper will expound on image captioning.

2. Methods

The two main categories of image captioning techniques are those relying on template methods and those based on encoder-decoder structures. Utilizing templates is mostly how the initial picture captioning method is implemented. This method first employs various classifiers, like SVM, to extract a number of important feature data, including such special attributes and key objects, and then transforms the data into descriptive text using a lexical framework or other targeted templates. The picture's object, the activity of the object, and the context within which the item is situated make up the majority of the important feature data. Some typical smoothing techniques eliminate the noise component. The item there in the picture is included in the feature data gleaned by Kulkarni et al. [1], and the connection between both the information extracted and the eventual image description outcome is assessed before choosing the optimum phrase arrangement to create the image description. Additionally, it employs the detectors to explicitly extract the properties of the object and the connections between those characteristics and the object, create a piece of tripartite information set using CRF to house the collected information, and then apply the predefined rules to build the summary. The aforementioned techniques rely far too heavily on extracted detailed information and predefined rules, and differing rules will have a direct impact on the output statements. The resulting image description information is overly straightforward and unadaptable, which frequently fails to get the required results. The encoder-decoder-based picture captioning paradigm offers greater prediction flexibility. Convolution neural networks are employed in the encoder to vectorize and extract the primary image feature data; recurrent neural networks are primarily utilized in the decoder to integrate the vectorized feature representations with semantic data to provide a descriptive assertion of the picture content [2].

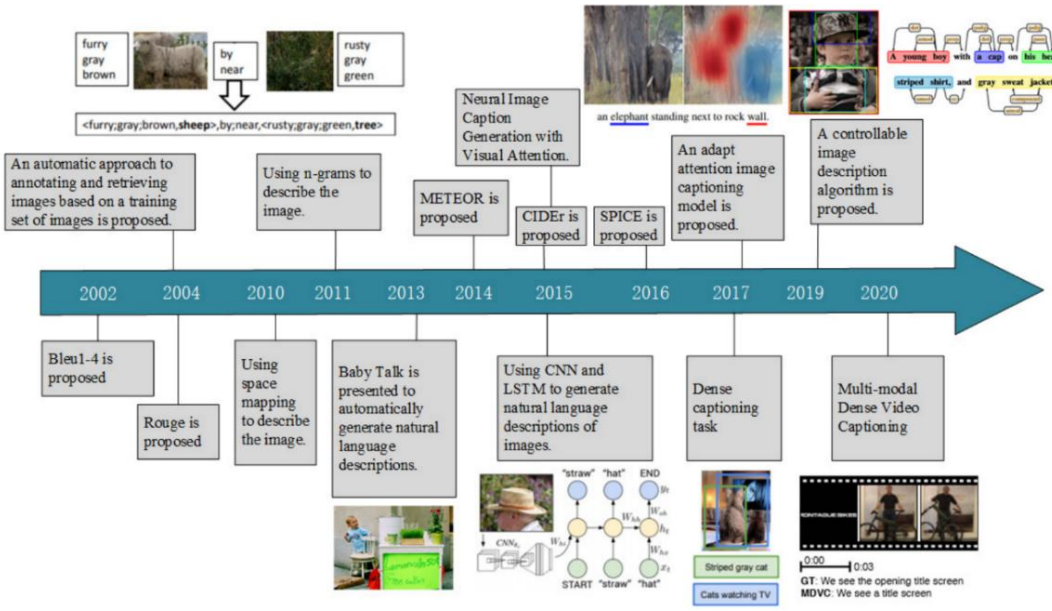

Figure 1. Development History [3].

DOI: 10.54254/2755-2721/5/20230511

Only the picture features are conveyed in the first stage of deciphering in the model, which is straightforward. The crucial picture feature data is lost as the forecast sequence's time step lengthens, making the outcome of image prediction less than ideal. Researchers tried using attention mechanisms to address the issue of image captioning as soon as they were discovered. They developed two strategies for attention and used them on various visual regions. Each area of the image receives a distinct weight value as a result of the soft learning algorithm. The region is much more crucial to prediction the greater the value. Notably, the weight distribution for every region is then utilized to determine the context vector that will be used. This hard overview of the system, which differs from the soft attention aspect in that it only concentrates on one area, is typically used to promote flexibility. The likelihood value of the spot you randomly picked is calculated. The top-down and bottom-up picture description modalities are combined there in the model structure model. Consequently, the network has the ability to concentrate attention on both significant semantics as well as important characteristics of images, judging how to do so flexibly and accurately.

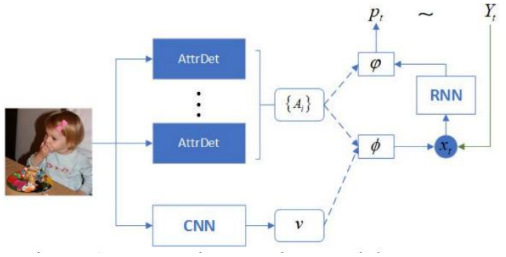

Figure 2. Model Structure of Semantic attention [3].

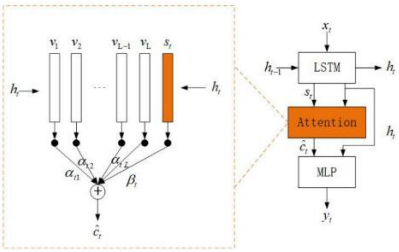

Figure 3. Model structure of Adaptive attention [3].

Primarily, the coevolutionary neural network is employed to process the feature information of the image, then the attribution detector is implemented to determine the visual attributes of objects, and ultimately the two are combined to produce the final picture description sequence that uses the recurrent network. When people interpret the image features, there are actually only two types of words included in the visual features: visual words, that may be successfully predicted utilizing image data, and non-visual words, which are words with the forms "and," "to," and "a," that cannot be acquired by assessing image data and must instead be implied from the contextual information. This model includes a visual guiding indication with some semantic information that is employed to assess if an attention mechanism is required given the circumstances. The multiregional characteristics acquired by image combination will be given varied ratios if the process is required.

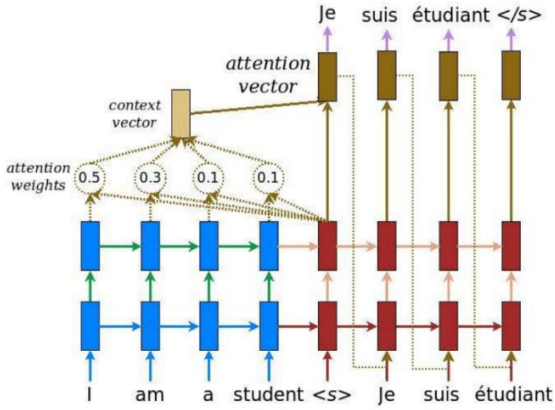

Figure 4. The mechanism of attention in machine translation [6].

DOI: 10.54254/2755-2721/5/20230511

Hossain et al. [4] study the availability of attentiveness in the vision and image system and present a picture caption model based on "attentiveness" as an excellent illustration of drawing inspiration from the human brain. According to this approach, attention refers to the capacity to dynamically highlight the critical component of a picture. In particular, Hossain et al. [4] demonstrate how well the model can suddenly adjust its focus to the essential item when creating the corresponding words. They develop two mechanisms—a "hard" deterministic attention mechanism and a "soft" deterministic attention mechanism—and train them using conventional back-propagation techniques while maximizing an approximation of the variational lower bound or something analogous [5]. This approach also has the benefit of roughly depicting what it "sees" to glean insights. A potential drawback is that focusing on the most prominent object would result in losing other, less critical information, leading to less detailed and plentiful captions.

We are all aware that creating an image caption involves two processes: comprehending the input and coming up with words. The second process also requires a language model. Many academics address this issue by independently resolving the two methods and then fusing them. This, however, is insufficient since the brain quickly converts a vast amount of visual data into descriptive language. A model is proposed concerning a propose a predictive model utilizing a deep reoccurring architecture after being impressed by the most recent advancement in translation software that recursive human brains (RNN) can complete the translation that typically requires a series of subtasks, and even in a more accurate and much simpler way [5]. Instead of using the decoder RNN, which is initially learned for a classification job, the deep neural network (CNN) is employed. As for the RNN decoder, the final hidden layer is used as the input, which produces the phrase. An end-to-end system, this neural network is wholly trainable and can be improved via stochastic gradient descent. The goal of this model is to maximize conditionally likelihood p(S|I), where S is the output phrase, and I is the input image [5]. They choose to employ a Long-Short Term Recollection Sentence Generator as their RNN alternative, frequently used for translation and creation activities.

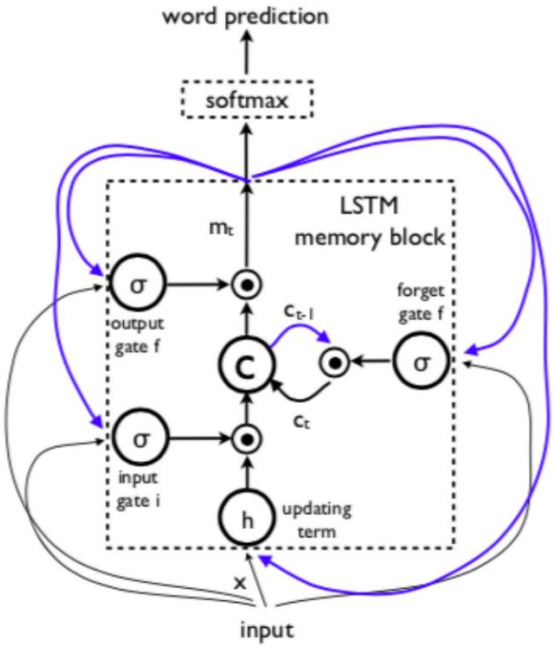

Figure 5. Memory block schematic diagram [6].

Along with these two successes, Kim et al. [7] presented a report in 2020 that was based on deeper mimicry of the human brain. When people understand sentences, they automatically form an image in their minds, which tends to last longer in the brain than the actual sentences, which inspired this work

DOI: 10.54254/2755-2721/5/20230511

[7]. In order to incorporate this into models, Kim and colleagues discussed the joint feature spaces of photos and their captions. They combine sentence and image elements into one common area, where new descriptions can be created from descriptions and unique descriptions can be made from images [7]. When a word is formed or read, the visual representation will be updated to incorporate the new information. The model may also capture vision scenes dynamically from the described images [8]. This procedure is similar to how ideas are stored in long-term memory. RNNs are employed in the essay to achieve this. The efficiency of bi-directional recovery is also looked at and contrasted in the model summary table, and the RNN baseline is also included in the article.

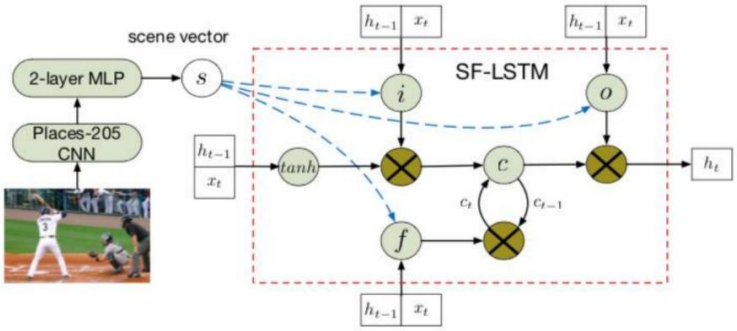

Figure 6. Schematic diagram of the RNN [6].

When the focus shifts between different regions of a picture during object recognition, a strand of visual order can be described by the sequence in which the observer's brain interprets the image's various abstract meanings. Accordingly, Kim et al. [7] generation model is based on the idea that phrases and pictures have a parallel structure. In particular, the paper assumes a close correlation between visual conceptions and literary realizations. The method used to produce new words from previously generated ones aligns with how people see images. The article uses deep recurrent neural RNNs to construct the model, encoding the semantically standard content between text descriptions of visual situations [9]. This network's hidden state is utilized to forecast the location of the subsequent visual focus and the following word in the associated text [9]. In the essay, another paradigm for scenario-specific contexts is also introduced. The model collects the image's high-level semantic information, such as the impact of the environment in which the photo was shot on the potential behavior of the subjects. The language model was trained to produce words consistent with the chosen scenario categories.

In order to do this, the scenario context gathers image feature matrices from the image, skewing the variables in the RNNs in the process. The available techniques for evaluating subtitles primarily evaluate the similarity of natural language descriptions by examining their sensitivity to n-gram overlap [8]. The main disadvantage of these methods is that they could give high ratings to words that just had the same format, and in many instances, this deviates from how people perceive an image. In light of this, Lee [10] suggests using the SPICE metric for new caption evaluation. This metric provides assessments by looking at the semantic content of the descriptions and a better understanding of how people evaluate photos. In addition, SPICE could examine any model's performance in greater detail than other computerized evaluation metrics [10]. For instance, it can determine if a model can count and which models can best depict the image's color. They link interdependence trees to scene graphs using a rule-based framework after applying a dependency parser to determine the semantic dependencies between words pre-trained on an extensive dataset [7]. SPICE calculates an F-score based on the combination of perfectly rational item sets expressing semantic assertions in the scene graph for the specified candidate and reference scenes.

Today, top-down mechanisms are the most prevalent ones that depend on visual attention. They receive their partially completed caption at each time step to provide context. The problem with these

DOI: 10.54254/2755-2721/5/20230511

models is that they do not consider what parts of a visual will be paid attention to [6]. These have a bearing on the captions' quality by emphasizing essential details. Regions of objects will offer similar descriptions to ones supplied by people. Present Up-Down is a model that combines a top-down context-specific process with a visual bottom-up mechanism [10]. The first offers suggestions for image regions. The former employs contextual to determine a concentration distribution over the items it determines to be salient. In contrast, the latter allows attention to be focused on the input's most crucial items image.

3. Related Works

A growing body of research on image captioning is available, and it may be broadly categorized into bottom-up and top-down approaches. The "traditional" methods use language models to construct sentences from the ground up, starting with visual objects, concepts, qualities, phrases, and words. The "contemporary" methods are top-down, which frame the issue of image captioning as a component of computational linguistics. These methods convert to a language equivalent rather than between several languages. It is noteworthy that a convolutional neural network is commonly pretrained for classification tasks on huge datasets, and provides visualization. RNN-based language models are used to translate. The key benefit of this strategy is that the whole network can be trained from start to finish, meaning that all the variables may be discovered through data. The type of recurrent networks utilized determines how the various methods differ from one another often. The most advanced method for solving this issue is top-down. Although visual attention has long been understood in psychology and neuroscience, machine learning and allied fields have only just begun to investigate it.

Deep learning-based algorithms for captioning images can produce captions of both visual and multimodal spaces. Naturally, databases for image captioning contain text versions of the associated captions. The input images and the related captions are separately provided to the language processor in the optical space-based approaches. Conversely, common multimodal space is acquired from the visuals and the related captions in the multimodal space example. The language decoder is subsequently given access to this multimodal description. Language encoder, multimodal space part, vision part, and language decoder are all components of a conventional multimodal space-based program's design. In order to gather the visual features, the vision component uses a deep neural network as a learning algorithm. The language encoder portion generates a dense feature encoding for every word after extracting the word aspects. It is noteworthy that the semantic, temporal background is subsequently transmitted to the recurrent levels. The multimodal space component places the word characteristics and image aspects in the same spatial location. A RNN for phrases and a convolutional network for visuals make up the m-RNN(multimodal Recurrent Neural Network) approach [11]. The entire m-RNN model is made up of two sub-networks and their multimodal layer of interaction. The input for this approach includes both images and sentence fragments. To produce the following word of captions, it computes the probability distribution. Throughout this model, there really are five additional layers: recurrent layer, SoftMax layer, multimodal layer, and two-word encoding layer. Considering that it is frequently impractical to annotate data precisely, unlabeled data are becoming more and more prevalent in our daily lives. As a result, for picture captioning, researchers are currently concentrating on more unsupervised learning-based and reinforcement learning algorithms. An individual using reinforcement learning makes a decision, obtains rewards, and then transitions to a separate state. The agent makes an effort to choose an action with the prospect of receiving the greatest possible long-term reward. To deliver the assurances of a functional form, it requires action information and a continuous state.

Conventional reinforcement learning techniques have a number of drawbacks, including unclear state data and the absence of assurances for a value function. Employing optimization approaches and gradient descent, a subset of reinforcement learning, policy gradient methods can select a certain policy for a given action. For action that ensures congruence, the policy may include domain knowledge. So, compared to systems based on value functions, policy gradient technique requires a number of parameters. To retrieve picture features, convolutional learning-based algorithms for image

DOI: 10.54254/2755-2721/5/20230511

captioning utilize numerous types of image encoders. Consequently, the language decoders built on neural networks are then supplied the features to produce captions. The two primary problems with the methods are that they are developed via back-propagation techniques and maximum likelihood estimation. In this instance, the image and all of the successfully made ground-truth words are used to forecast the following word. As a result, the produced captions appear to be accurate. The term "exposure bias problem" refers to this phenomenon. Additionally, test-time assessment measures are non-differentiable. For instance, In a perfect world, exposure bias should really be avoided while training sequence models for picture captioning, and parameters should be explicitly optimized for the testing time. The operator can be trained by using the critic to estimate the anticipated reward in the long term. Image captioning methodology based on reinforcement learning extracts the following token from the system according to the benefits they obtain in each stage. Reinforcement learning techniques that use policy gradients can maximize the gradient and forecast the long-term cumulative rewards. It can therefore address the non-differentiable issue with parameters.

By combining a convolutional network, LSTM language, and a dense localization layer, dense captioning suggests a deep convolution localization network. Along with a single, effective forward pass, the compact localization layer efficiently analyses a picture and explicitly foretells a number of regions of interest. Therefore, unlike Fast R-CNN, it does not require external object proposals. The Faster R-work CNN is connected to the localization layer's operating theory [12]. Rather than using an ROI pooling method, Anderson et al. [13] adopt a spatial, differential soft attention mechanism with smoothing. This change makes it easier for the approach to identify the active areas and backpropagate over the network. For the trials, Visual Genome dataset is used in creating region-level image captions. One subjective explanation of the whole visual scene is insufficient to convey the full idea. Global image descriptions lack the objectivity and level of detail of region-based explanations. Nonetheless, dense captioning refers to the location-based description. The intensive captioning process presents various difficulties. One object could contain several overlapping zones of interest due to how densely packed the regions are. Additionally, it is exceedingly challenging to identify each region of interest for all visual conceptions. Another dense captioning technique was put out by Hossain et al. [4]. These problems can be solved using this approach. First, it handles an inference technique that depends on both the region's visual characteristics and its anticipated captions. This enables the model to determine where the bounding box should be placed. Second, they use a context fusion technique to create a comprehensive semantic description by fusing context features together with local visual features. A sizable number of techniques have so far generated image captions in a suitable manner. Testing samples and training are drawn from the same realm for the procedures. As a result, it is uncertain if these strategies will work effectively with open-domain photos. Additionally, they are only proficient at identifying generic visual material. Some important objects, including famous people and famous places, fall outside of their purview. These algorithms generate captions that are assessed using automated metrics. Various evaluation metrics have already produced promising outcomes for these techniques. The human perception of evaluation and the assessment of metrics, however, differ substantially in terms of performance.

The authors provided a thorough analysis of the most recent deep learning-based image captioning methods by late 2018, as was noted in the review publication [7]. The research topic was covered from various angles, including gaining knowledge, design, amount of captions, communicative language teaching, and feature mapping. The report provided a taxonomy of the available methodologies and contrasted their advantages and disadvantages. The advantages and disadvantages of different data sets and evaluation methods are also discussed. Another review study (Lee [10]) on color-based segmentation, N-cut, and hybrid engines was released in the middle of 2019. The best accuracy for these models is achieved by model engineering, which involves adding more hyper-parameters to the pipeline. Another study (Chenyu [6]) from the same year reviewed the literature from 2017 to 2019 and addressed various datasets and architectures. They claimed that CNN-RNN models are beaten by the CNN-LSTM ones, with BLEU being the most used evaluation metric (1 to 4). They also discovered that the attention mechanism and encode decode work best for implementing such a model.

DOI: 10.54254/2755-2721/5/20230511

Additionally, they claimed that combining the two approaches can help to improve the outcomes of such a task. The research field of image captioning is still active, and new techniques are still being published today [10]. One of the primary objectives of writing this review study is to address all of the most current advancements in the last several years, including 2020.

On two datasets, Flickr30k and MSCOCO, the authors analysed image captioning techniques from 2016 to 2019, including more recent ones. They have tested different extractors, such as GoogleNet with all nine Inception models, VGG-16 Net, ResNet, AlexNet, and DenseNet [3]. Language models were also discussed, including TPGN, GRU, CNN, RNN, and LSTM. This comparison involves the evaluation measures BLEU (1–4), CIDEr, and METEOR. (Wang et al. [3]) highlighted some improvements made to the image captioning assignment until the beginning of 2020, when several methods were discussed.

4. Datasets and Evaluation Indexes

We need data sets as input in order to construct a superior network model. We can prevent overfitting phenomena by studying a plenty of data sets, and the accuracy of the network model can also be improved. We must utilize the assessment index to assess a model's quality. Because each evaluation index has a numerical value, we can see the model's benefits and drawbacks. Here, the typical datasets and several assessment indices for the method for captioning images are introduced.

4.1. Dataset

The data sets for images and their accompanying descriptions must both be included in image captioning algorithms. MSCOCO 2014, flickr30k, flickr8k, ICC, and Visual Genome are the key data sets connected to this kind [6]. Michael Bernstein created the image-intensive annotation data set known as Visual Genome, which includes 108,077 visual information. About 35 things are shown in each photograph, and each one is labeled. With a total of 540,000 domain descriptions, each image has 50 regional descriptions. There are also 280,000 picture attributes and 170,000 QA pairings. 230,000 image relations exist [6]. The AIC dataset contains the ICC dataset, which provides descriptions of Chinese image data. Most of the 300,000 photos are there. Each image has 1,500,000 Chinese description statements overall, with five statements per image. Flickr has gathered data sets called Flickr8k and Flickr30k that describe various human activities [6]. Most of the 8,000 photographs in Flickr8k are set as the training data, with the remaining 1,000 being used for verification and the final 1,000 for testing. 31,000 photographs, including 29,000 training datasets, 1,000 verification, and test sets, make up most of Flickr30k. Microsoft created the MSCOCO 2014 picture data set, which can be used for object recognition, semantic segmentation, and image captioning. There are numerous publications of the data set. Mscoco 2014 includes 40,504 verification samples, 40,775 test samples, and 82,783 training samples. There are also 886,000 segmented object photos and 270,000 segmented portrait images [6]. The MSCOCO 2014 dataset is frequently utilized in image captioning research because it contains more instances of annotated image captions than other data sets

4.2. Evaluation Indexed

The primary categories used to evaluate picture captioning are BLEU 1-4, ROUGE, METEOR, SPICE, and CIDEr. BLEU was initially employed to assess the caliber of translation output. It primarily consists of three components the accuracy degree of the coincidental n-gram ratio; the compensation mechanism, which mainly addresses the issue of too little length in the forecast sequence; and the geometrical average, which primarily balances the variance in the n-gram accuracy measures [1]. The ROUGE approach was created to assess how well the article contains a summary performed. ROUGE primarily consists of the four assessment indices ROUGE N, ROUGE S, ROUGE W, and ROUGE L [1]. The effect of picture labeling prediction can be assessed using the latter three techniques. METEOR is another evaluation metric used to gauge the impact of translation. A new penalty mechanism that METEOR utilizes is primarily used to punish situations in which the expected sequence does not match the order of the tag words [1].

DOI: 10.54254/2755-2721/5/20230511

While at the same time, METEOR replaces the unigram employed in Bleu with chunks, each comprising two neighboring unigram words. By using SPICE, the overlapping issue brought on by n-grams is avoided. Prior to mapping, it uses a dependency parser to encode the anticipated sequence and the accurate tag into a semantic dependency tree [1]. The map can be separated into tuples that contain objects, relationships, and attributes after mapping. The evaluation findings are acquired by comparing the graph created from the expected sequence with the map created from the actual label. Sentence matching is not how CIDEr assesses the efficiency of forecast description; instead, it looks at how closely the predicted text matches the actual label. The anticipated description is more appropriate, and the forecast effect is better the closer the two are in terms of resemblance. This index is now utilized more frequently than the previous evaluation metrics for assessing the quality of image descriptions.

5. Conclusion

The concepts of specific popular picture captioning techniques are introduced and examined in this work. Even though the forecast impact of the currently available image captioning algorithms has increased to some level, they do not fully achieve the function of creating particular summary comments in accordance with certain conditions. Necessary datasets and assessment indices for this field were introduced.

With the development of technology, image captioning could be enhanced from three perspectives. The first is to make the network more adaptable for the model to focus on particular situations and issues and produce tailored explanations following various circumstances and concerns. The second part involves improving the evaluation algorithm to assess the efficiency of the model's outcome sequence. The third goal is to make the model more robust and prevent interfering qualities from impacting the model's outcome.

References

[1] Kulkarni, G., Premraj, V., Ordonez, V., Dhar, S., Li, S., Choi, Y., & Berg, T. L. (2013). Babytalk: Understanding and generating simple image descriptions. IEEE Transactions on Pattern Analysis and Machine Intelligence, 35(12): 2891-2903.

[2] Herdade, S., Kappeler, A., Boakye, K., & Soares, J. (2019). Image captioning: Transforming objects into words. Advances in Neural Information Processing Systems, 32.

[3] Wang, C., Zhou, Z., & Xu, L. (2021). An integrative review of image captioning research. In the journal of physics: conference series (Vol. 1748, No. 4, p. 042060). IOP Publishing.

[4] Hossain, M. Z., Sohel, F., Shiratuddin, M. F., & Laga, H. (2019). A comprehensive survey of deep learning for image captioning. ACM Computing Surveys (CsUR), 51(6), 1-36.

[5] Elhagry, A., & Kadaoui, K. (2021). A thorough review of recent deep learning methodologies for image captioning. arXiv preprint arXiv:2107.13114

[6] Chenyu, C. (2020). Understanding Image Caption Algorithms: A Review. In Journal of Physics: Conference Series (Vol. 1438, No. 1, p. 012025). IOP Publishing.

[7] Kim, H., Tang, Z., & Bansal, M. (2020). Dense-Caption Matching and Frame-Selection Gating for Temporal Localization in VideoQA. arXiv preprint arXiv:2005.06409.

[8] Staniūtė, R., & Šešok, D. (2019). A systematic literature review on image captioning. Applied Sciences, 9(10), 2024.

[9] 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

[10] Lee, S., & Kim, I. (2018). Multimodal feature learning for video captioning. Mathematical Problems in Engineering, 2018.

[11] Rennie, S. J., Marcheret, E., Mroueh, Y., Ross, J., & Goel, V. (2017). Self-critical sequence training for image captioning. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 7008-7024).

DOI: 10.54254/2755-2721/5/20230511

[12] Yao, T., Pan, Y., Li, Y., & Mei, T. (2018). Exploring visual relationship for image captioning. In Proceedings of the European conference on computer vision (ECCV) (pp. 684-699).

[13] Anderson, P., Fernando, B., Johnson, M., & Gould, S. (2016, October). Spice: Semantic propositional image caption evaluation. In European conference on computer vision (pp. 382-398). Springer, Cham