DOI: 10.54254/2755-2721/8/20230196

Deep learning based automated image captioning system

S.P. Siddique Ibrahim[1, 6], S. Selva Kumar[2], C. Bharathi Priya[3], S. Vinoth Kumar[4], P. Parthasarathi[5]

[1,2]Assistant Professor, School of Computer Science and Engineering, VIT-AP University, Beside AP Secretariat, Amaravati. Andhra Pradesh, India
[3]Assistant Professor, Department of Computer Science and Engineering, Kumaraguru College of Technology, Coimbatore.
[4]Associate Professor, Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Chennai.
[5]Assistant Professor, Department of Computer Science and Engineering, Bannari Amman Institute of Technology, Sathyamangalam, Tamil Nadu, India.

[6]siddique.ibrahim@vitap.ac.in

Abstract. With the rise of deep learning in recent years, the merging of computer vision and natural language processing has inspired a huge interest. Image captioning is a unique deep learning application that has advanced rapidly in recent years. The ability to train a computer to properly describe the content of an image or an environment like a person has huge implications in computer vision, economics, and a variety of other fields. Across the year, this has been a difficult issue in the field of artificial intelligence, and many researchers have made tremendous progress. In this paper, we describe various deep neural network-based picture caption generation models, including RNN-based decoding and CNN-based accessing and Reinforcement-based framework. We also created captions for some of the images and compared the various feature extraction and encoder models to see which one provides the best accuracy and produces the desired outcome. Extensive experiments on the dataset reveal that the proposed framework outperforms the existing encoder based approach.

Keywords: deep learning, S-commerce, sentiment analysis, accuracy, specificity, sensitivity.

1. Introduction

Computer vision has made substantial advances in the image processing sector in recent years, such as image segmentation [1] and object detection [2]. Automatic visual captioning is a difficult problem with many complications that has been the subject of several significant academic studies. A broad range of applications exist for automatically creating complete and natural picture descriptions, including news photo labels, medical image descriptions, text-based image retrieval, information for blind users, and human-robot interaction. These image inscribing applications are helpful for both hypothetical and down to earth research. As a result, picture captioning has become a more difficult yet important activity in the age of artificial intelligence.

Given a visual picture, a picture inscribing calculation ought to create a semantic portrayal of the picture. The information picture in Figure 1 for instance, has a cow, grass, and water. An expression at the lower part of the page depicts the picture's substance, including the things that show up in the picture, the activity, and the scene.

People can without much of a stretch comprehend picture content and express it in regular language sentences as indicated by unambiguous requirements for picture subtitling; nonetheless, PCs require the coordinated utilization of picture handling, PC vision, normal language handling, and other significant areas of exploration results for picture inscribing. The objective of picture inscribing is to make a model that can completely take advantage of picture information to give more human-like rich picture inscriptions. Numerous strategies for picture inscribing have been proposed, including object distinguishing proof models, visual consideration-based picture subtitling, and Image Captioning utilizing Deep Learning. There are a few profound learning models, for example, the Inception model, VGG model, ResNet-LSTM model, and exemplary CNNRNN model in Deep Learning. [3,4].

The possibility to dissect their states, figure out their relations, and produce a semantically and linguistically accurate statement is expected for the significant description generation cycle of undeniable level picture semantics. Traffic data analysis [5], Self-driving cars, human-computer interaction [6], recommendation systems, medical image captioning, and automatic medical prescription [7], medical assistance, quality control in industry [8], for social media, and assistive technologies for hearing, handicap and visually impaired persons [9] are some of the applications of automatic image captioning. Given the numerous issues and barriers that come with surviving with a visual impairment, finding a way to reduce these issues may be extremely beneficial to these people and enhance their quality of life. For picture captioning, this article used a CNN-based and reinforcement-based approach [10].

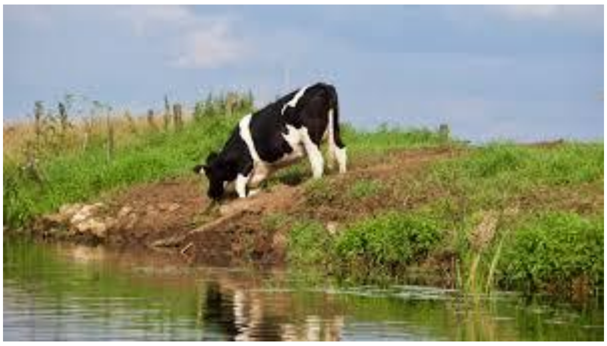

Figure 1. An example of automatic image caption - cow eating grass near body of water.

2. Literature review

Template based strategies that utilized image order, for example allocating marks to objects from a foreordained arrangement of classes and placing them into an sample layout express, were among the endeavours to settle this test. Encoding is done with Convolutional Neural Networks, while decoding is done with Recurrent Neural Networks [11]. The pictures are transformed to vectors using CNN, and these vectors are referred to as image features, and they are then fed into recurrent neural networks as input. The actual captions for the project are obtained using RNN's NLTK library. Only CNN is employed for picture encoding and decoding in the CNN-CNN based framework [12]. To acquire the result, a vocab dictionary is employed and it is mapped with Image characteristics [13]. As a result, the caption is free of errors. The train, which is composed of several models that are supplied at the same time of convolution methods concurrently, is unquestionably faster than the train, which is composed of a continuous flow of recurrently repeat of these techniques. In comparison to the CNN-RNN Model, the CNN-CNN Model requires less training time. The CNN-RNN Model takes longer to train since it is sequential, but it loses less data than the CNN-CNN Model. However, most recent research has concentrated on Recurrent Neural Networks [14]. RNNs are already widely used in a variety of Natural Language Processing activities, such as machine translation, which generates a series of words. The image caption generator adds to the same programme by creating a word-by-word explanation for a picture. The parallax mistake is one of several hurdles and unresolved problems in picture captioning that are inherent in the problem's nature. It could in fact be challenging for the human eye to see a thing when it is seen from certain angles that make the item's appearance modify to where it is as of now not perceivable. Numerous things of different structures, notwithstanding points, may have a place with a given article class. Things that are covered by different articles could likewise make it harder for the visual aide to identify every one of them suitably. Object recognition is additionally hampered by scene mess. While fostering a visual aide fit for subtitling pictures, it's critical to perceive and resolve these issues.

3. Re-enforcements based framework

Algorithms for image captioning are often classified into three types. The first way is depicted in Figure 2. Handle this issue using a retrieval-based technique, which first fetches the closest matched photos before displaying the caption. Instead of dealing with fresh images, this technique works with grammatically acceptable statements.

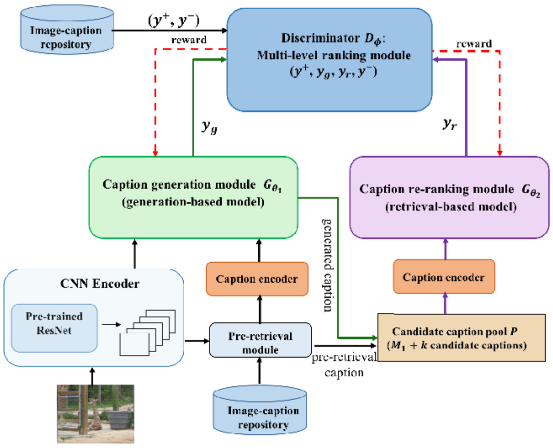

Figure 2. Image caption based on retrieval.

The second technique, shown in Figure 3, uses a template to divide the caption image sentence into separate portions based on predetermined rules. These approaches utilize a rigorous sentence structure to generate a complete sentence after using multiple classifiers to detect the objects in a picture, as well as their features and relationships. Despite the fact that these approaches can produce a new phrase, they are unable to accurately convey the visual context or construct flexible and meaningful words. The third technique, which is based on neural networks and inspired by encoder and decoder methods, is used in modern picture caption research.

3.1. Problem formation

We figure out image captioning as a decision-making process. In navigation, there is a specialist that communicates with the climate, and executes a progression of activities, to improve an objective. In picture subtitling, the objective is, given a picture M, to create a sentence S ={ S1,S2,S3… .Sn} which accurately portrays the picture content, where Si is a word in sentence S and L is the length. Our model, including the strategy network Pθ and esteem network vθ, can be seen as the specialist; the climate is the given picture M and the words anticipated so far {s1,s2… sn} and an activity is to foresee the following word st+1.

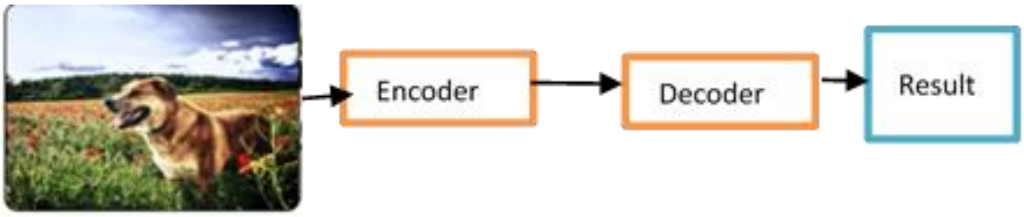

Figure 3. Image caption based on neural network.

3.2. Deep inforcement learning

Because of its unsupervised nature, reinforcement learning methods are widely used in a variety of applications, from gaming to Facebook, as shown in Figure 4. For picture captioning, there has been a recent trend toward less reliance on paired image-caption databases. Gu et al. [15] accomplished some early work that involved making inscriptions in a turn language and making an interpretation of them to an objective language. For the turn language, this method requires a matched picture subtitle dataset, yet it doesn't need a coordinated dataset with inscriptions in the objective language. Reinforcement learning using gradient policy and RNNs was employed in another research work on this topic in 2016 [9]. Oriol et al. [11] were the first to propose research that looked at employing conditional GANs to produce human-like and diversified descriptions. Despite the fact that this strategy improves caption variety, it compromises overall performance and resulting in poorer language metric scores as compared to other methods. For unsupervised training, shel et al. [9] utilize a bunch of pictures, a text corpus, and a visual idea locator. The visuals and sentence corpus are projected into a similar idle region with the goal that they might be recreated. The captions from Aluli [13], a photo-sharing company, were used to create the sentence corpus. Each photograph on this platform is accompanied with a caption. This corpus is separate from the pictures and is unrelated to them.

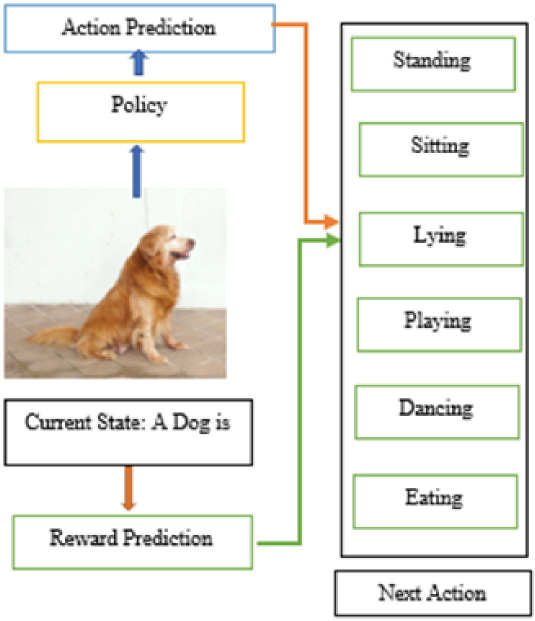

Figure 4. Proposed Reinforcement learning framework.

A picture encoder, an expression generator, and a discriminator attempt to repay the proposed structure. Since no picture inscription matches exist, three new measurements have been acquainted as three discriminators with evaluate the model's exhibition. At each time step, the discriminator recognizes between a genuine sentence from the sentence corpus and a sentence made by the model, and the generator is compensated. The generator plans to produce credible explanations by augmenting this award. Be that as it may, this discriminator isn't enough since the nature of the created sentence and its pertinence to the picture should likewise be assessed. To do as such, the model must initially grasp the picture's visual items. The made words are granted assuming that they contain words whose proper visual thought is perceived inside the picture. A "thought reward" is a kind of remuneration. At long last, on the grounds that the model's presentation is profoundly reliant upon the visual idea locator's exhibition, and these identifiers can recognize a predetermined number of things, pictures and subtitles are projected onto a common dormant space so they might be recreated.

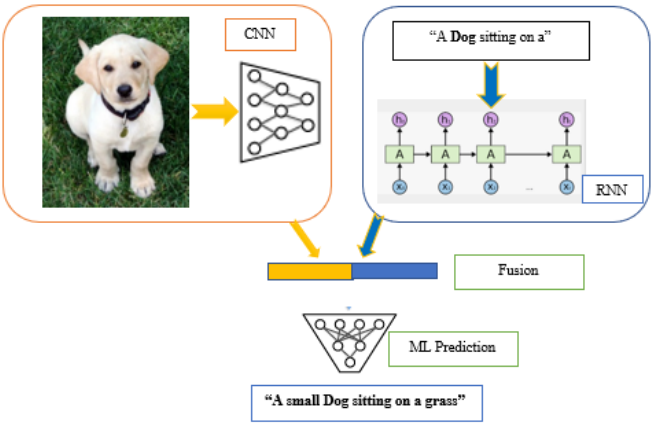

Figure 5. Illustration of proposed system.

This chapter present a unique decision-making framework for image captioning in this research. We use a "policy network" and a "value network" to jointly select the next best word at each time step, rather than learning a sequential recurrent model to aggressively search for the next accurate sentence presented in Figure 5. As a local agent, the policy network gives the confidence of predicting the following word based on the current state. The value network acts as a global and lookahead guide, evaluating the reward value of all conceivable extensions of the present state. The objective of such a value network shifts from predicting the correct words to creating captions that are comparable to ground truth captions. By employing the policy network alone, our system is able to incorporate good phrases that have a low chance of being drawn. The suggested automated cation based on reinforcement learning is demonstrated in Figure 5. At this time, our policy network's top choice is not "holding." However, our value network advances one step to the stage where holding is formed and assesses how good that state is for the purpose of creating a nice caption in the end, as shown in Fig.6. Both networks are able to pick the word holding and complement one another. We employ deep reinforcement learning with embedded reward to train the policy and value networks. We start by pretraining a policy network with cross entropy loss and a value network with mean squared loss using ordinary supervised learning. Then, using deep reinforcement learning, we strengthen the policy and value networks. The proposed method used public MS COCO dataset for automatic image captioning. For fair comparison we used ten-fold splits used in [7,10]. The system uses 87,500 images for training and 5,000 images for testing purpose. The desktop machine with AMD Ryzen with 16GB RAM is used for implementation. The proposed work's caption technique compares with existing work depicted in Figure 6.

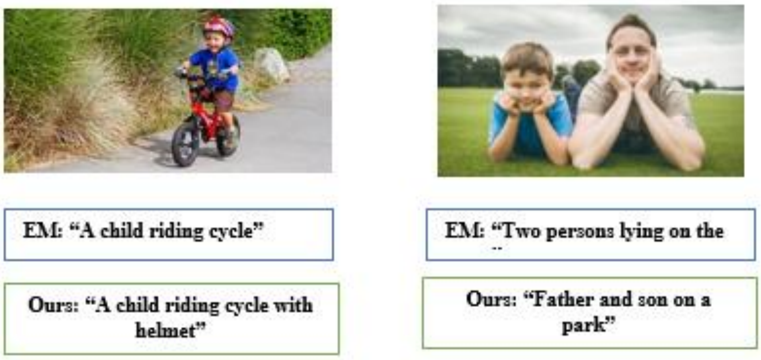
EM: "A child riding cycle"
Ours: "A child riding cycle with helmet"
EM: "Two persons lying on the
Ours: "Father and son on a park"

Figure 6. llustration of results of our method and conventional existing method (EM). In the pictures our method generates better results.

4. Conclusion

We introduce a unique decision-making system for automatic picture captioning in this paper, which produces state-of-the-art results on conventional benchmarks. In contrast to earlier encoder-decoder frameworks, our technique generates captions using a value network and a policy network. Our suggested model is not ideal for all of the testing examples, and it may occasionally produce inaccurate captions. As a result, we'll be working on upgrading network topologies and studying incentive design by taking into account various embedding measures in the future.

References

[1] Easwaramoorthy, S., Moorthy, U., Kumar, C. A., Bhushan, S. B., & Sadagopan, V. (2018). Content based image retrieval with enhanced privacy in cloud using apache spark. In Data Science Analytics and Applications: First International Conference, DaSAA 2017, Chennai, India, January 4-6, 2017, Revised Selected Papers 1 (pp. 114-128). Springer Singapore.

[2] Subramanian, M., Cho, J., Sathishkumar, V. E., & Naren, O. S. (2023). Multiple Types of Cancer Classification Using CT/MRI Images Based on Learning Without Forgetting Powered Deep Learning Models. IEEE Access, 11, 10336-10354.

[3] Sathishkumar, V. E., Cho, J., Subramanian, M., & Naren, O. S. (2023). Forest fire and smoke detection using deep learning-based learning without forgetting. Fire Ecology, 19(1), 1-17.

[4] Kogilavani, S. V., Sathishkumar, V. E., & Subramanian, M. (2022, May). AI Powered COVID-19 Detection System using Non-Contact Sensing Technology and Deep Learning Techniques. In 2022 18th International Conference on Distributed Computing in Sensor Systems (DCOSS) (pp. 400-403). IEEE.

[5] Karpathy, Andrej, and F. F. Li. "Deep visual-semantic alignments for generating image descriptions." Computer Vision and Pattern Recognition IEEE, 3128-3137, 2015.

[6] Show and Tell: A Neural Image Caption Generator by Oriol Vinyal, Alexander Toshev, Samy Bengio, Dumitru Erhan, IEEE, 2015.

[7] Jagadeesh Kumar, S.J.K., Parthasarathi, P., Mehedi Masud, Jehad F. Al-Amri and Mohamed Abouhawwash, Butterfly Optimized Feature Selection with Fuzzy C-Means Classifier for Thyroid Prediction, Intelligent Automation & Soft Computing, Vol. 35, No.3, pp.2909–2924, 2023.

[8] Liu, Shuang & Bai, Liang & Hu, Yanli & Wang, Haoran.. Image Captioning Based on Deep Neural Networks. MATEC Web of Conferences, 2018.

[9] Subramanian, M., Rajasekar, V., VE, S., Shanmugavadivel, K., & Nandhini, P. S. (2022). Effectiveness of Decentralized Federated Learning Algorithms in Healthcare: A Case Study on Cancer Classification. Electronics, 11(24), 4117.

[10] Sathishkumar, V. E., Park, J., & Cho, Y. (2020). Using data mining techniques for bike sharing demand prediction in metropolitan city. Computer Communications, 153, 353-366.

[11] Vinyals, Oriol, et al. "Show and tell: A neural image caption generator." IEEE Conference on Computer Vision and Pattern Recognition IEEE Computer Society, 3156-3164, 2015

[12] Shanmugavadivel, K., Sathishkumar, V. E., Raja, S., Lingaiah, T. B., Neelakandan, S., & Subramanian, M. (2022). Deep learning based sentiment analysis and offensive language identification on multilingual code-mixed data. Scientific Reports, 12(1), 21557.

[13] Sathishkumar, V. E., Agrawal, P., Park, J., & Cho, Y. (2020, April). Bike sharing demand prediction using multiheaded convolution neural networks. In Basic & Clinical Pharmacology & Toxicology (Vol. 126, pp. 264-265). 111 RIVER ST, HOBOKEN 07030-5774, NJ USA: WILEY.

[14] Sathishkumar, V. E., Sharmila, C., Santhiya, S., Poongundran, M., Sanjeeth, S., & Pranesh, S. (2023, March). Convolutional Neural Networks for Traffic Sign Classification Using Enhanced Colours. In Deep Sciences for Computing and Communications: First International Conference, IconDeepCom 2022, Chennai, India, March 17–18, 2022, Revised Selected Papers (pp. 34-43). Cham: Springer Nature Switzerland.