Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 11 Issue 02, February-2022

Image Captioning System using Recurrent Neural Network-LSTM

Alok Mathur
B.Tech CSE Student, VIT, Vellore
Pune, India

***Abstract*— Image captioning may benefit the area of retrieval, by allowing us to sort and request pictorial or image-based content in new ways. Mapping the space between images and language, may resonate with some deeper vein of progress which, once unearthed, could potentially lead to more sophisticated machines. Generating captions for images is a vital task relevant to the area of both Computer Vision and Natural Language Processing. The main challenge of this task is to capture how objects relate to each other in the image and to express them in a natural language (like English). Image captioning has a variety of uses, including editing software recommendations, virtual assistants, image indexing, accessibility for visually impaired people, social media, and other natural language processing applications. We aim to achieve this task by employing a model based on recent advances in computer vision and machine translation.**

Keywords—Image Captioning, Deep Learning, Neural Network

I. INTRODUCTION

Image captioning is a branch of artificial intelligence research that focuses on recognising natural scenes and explaining them in natural language. Image captioning could help with retrieval by allowing us to organise and request pictorial or image-based content in new ways. Mapping the space between visuals and text could be a sign of deeper progress. Creating captions for images is an important task in both the fields of computer vision and natural language processing. The major goal of this work is to capture how items in the image interact with one another and represent them in plain language (like English).

To create lexically rich text descriptions for photos, neural networks can be employed. We've added an attention mechanism that can recognise what a word in an image refers to. This will be an effective tool for utilising large amounts of unformatted image data.

II. LITERATURE REVIEW

[1] Image Captioning based on Deep Reinforcement Learning

This paper proposes a model in contrast to existing sequential models such as Recurrent Neural Networks (RNNs). It utilizes two networks called ‘policy network’ and ‘value network’ to collaboratively generate the captions of images. Based on reinforcement learning, the policy network gives some possible actions about the object. Then, the value network makes decisions on whether to choose the action given by the policy network based on the evaluated reward scores. The policy network consists of two networks, a Convolutional Neural Network (CNN) and a Recurrent Neural Network (RNN), which provides a probability for the agent to take actions at each state. The value 9 network contains three parts, a CNN, a RNN and a Linear Mapping Layer, in which a perceptron model is utilized to evaluate the predictions to choose the most suitable *Image Captioning based on Deep Reinforcement Learning*

[2] Image-Text Surgery: Efficient Concept Learning in Image Captioning by Generating Pseudopairs

Image captioning aims to generate natural language sentences to describe the salient parts of a given image. Although neural networks have recently achieved promising results, a key problem is that they can only describe concepts seen in the training image- sentence pairs. Efficient learning of novel concepts has thus been a topic of recent interest to alleviate the expensive manpower of labeling data. Image-Text Surgery, to synthesize pseudo image-sentence pairs is proposed. The pseudo pairs are generated under the guidance of a knowledge base, with syntax from a seed data set (i.e., MSCOCO) and visual information from an existing large-scale image base (i.e., ImageNet). Via pseudo data, the captioning model learns novel concepts without any corresponding human-labelled pairs. adaptive visual replacement is introduced, which adaptively filters unnecessary visual features in pseudo data with an attention mechanism. approach on a held-out subset of the MSCOCO data set is evaluated. The experimental results demonstrate that the proposed approach provides significant performance improvements over state-of-the-art methods in terms of F1 score and sentence quality. An ablation study and the qualitative results further validate the effectiveness of the approach.

[3] Boosting Image Captioning with Attributes

Automatically describing an image with a natural language has been an emerging challenge in both fields of computer vision and natural language processing. it presents Long Short Term Memory with Attributes (LSTM-A) - a novel architecture that integrates attributes into the successful Convolutional Neural Networks (CNNs) plus Recurrent Neural Networks (RNNs) image captioning framework, by training them in an end-to-end manner. Particularly, the learning of attributes is strengthened by integrating inter-attribute correlations into Multiple Instance Learning (MIL). To incorporate attributes into 13 captioning, it constructs variants of architectures by feeding image representations and attributes into RNNs in different ways to explore the mutual but also fuzzy

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 11 Issue 02, February-2022

relationship between them. Extensive experiments are conducted on COCO image captioning dataset and their framework shows clear improvements when compared to state-of-the-art deep models. More remarkably, it obtains METEOR/CIDEr-D of 25.5%/100.2% on testing data of widely used and publicly available splits in when extracting image representations by GoogleNet and achieve superior performance on COCO captioning Leaderboard.

[4] Image Captioning with Scene-graph Based Semantic Concepts

Different from existing approaches for image captioning, in this paper, they explore the cooccurrence dependency of high-level semantic concepts and propose a novel method with scene-graph based semantic representation for image captioning. To embed scene graph as an intermediate state, the task of image captioning into two phases, called concept cognition and sentence construction respectively. a vocabulary of semantic concepts is built and propose a CNN-RNN-SVM framework to generate the scene- graph-based sequence, which is then transformed into a bit vector, as the input of RNN in the next phase. They evaluate their method on FLICR dataset. Experimental results show that the approaches obtain a competitive or superior result to the state-of the-arts.

[5] Image Captioning with Word Level Attention

Image captioning is an attractive and challenging task to perform automatic image description and a number of works are designed for this task. these researches are based on convolutional neural network (CNN) and recurrent neural network (RNN), where the primary input to language model for word prediction at the current time step is usually the linguistic word generated at the previous time step. In this work, a novel word level attention layer is designed to process image features with two modules for accurate word prediction. The first is a bidirectional spatial embedding module to handle feature maps, 14 then the second module employs attention mechanism to extract word level attention which will be fed into language model. The experimental results on the benchmark MSCOCO dataset demonstrate that the proposed model achieves the state-of-the-art performances with 106.0 on CIDEr and 34.0 on B-4.

[6] A parallel-fusion RNN-LSTM architecture for image caption generation

The models based on deep convolutional networks and recurrent neural networks have dominated in recent image caption generation tasks. Performance and complexity are still eternal topic. By combining the advantages of simple RNN and LSTM, a novel parallel fusion RNN-LSTM architecture is presented, which obtains better results than a dominated one and improves the efficiency as well. The proposed approach divides the hidden units of RNN into several same-size parts, and lets them work in parallel. Then, their outputs are merged with corresponding ratios to generate final results. Moreover, these units can be different types of RNNs, for instance, a simple RNN and a LSTM. By training normally using Neural Talk platform on Flickr8k dataset, without additional training data, better results are obtained than that of dominated structure and particularly, the proposed model surpass Google NIC in image caption generation.

[7] Image Captioning With Visual-Semantic Double Attention

A novel Visual-Semantic Double Attention (VSDA) model is proposed for image captioning. In their project, VSDA consists of two parts: a modified visual attention model is used to extract sub-region image features, then a new Semantic Attention (SEA) model is proposed to distill semantic features. Traditional attribute-based models always neglect the distinctive importance of each attribute word and fuse all of them into recurrent neural networks, resulting in abundant irrelevant semantic features. In contrast, at each 16 timestep, their model selects the most relevant word that aligns with current context. In other words, the real power of VSDA lies in the ability of not only leveraging semantic features but also eliminating the influence of irrelevant attribute words to make the semantic guidance more precise. Furthermore, their approach solves the problem that visual attention models cannot boost generating non-visual words. Considering that visual and semantic features are complementary to each other, their model can leverage both of them to strengthen the generations of visual and non-visual words. Extensive experiments are conducted on famous datasets: MS COCO and Flickr30k. The results show that VSDA outperforms other methods and achieves promising performance.

[8] Image Captioning with Affective Guiding and Selective Attention

Image captioning is an increasingly important problem associated with artificial intelligence, computer vision, and natural language processing. In this article, an image captioning model with Affective Guiding and Selective Attention Mechanism named AGSAM is proposed. In this approach it is aimed to bridge the affective gap between image captioning and the emotional response elicited by the image. First, an effective component that captures higher-level concepts encoded in images into AG-SAM is introduced. Hence, this proposed language model can be adapted to generate sentences that are more passionate and emotive. In addition, a selective gate acting on the attention mechanism controls the degree of how much visual information AG-SAM needs. Experimental results have shown that this model outperforms most existing methods, clearly reflecting an association between images and emotional components that is usually ignored in existing works.

[9] Image Captioning using Deep Neural Architectures

Automatically creating the description of an image using any natural language sentences is a very challenging task. It requires expertise in both image processing as well as natural language processing. This paper discusses different available models for image captioning tasks. The

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 11 Issue 02, February-2022

advancement in the task of object recognition and machine translation has greatly improved the performance of image captioning models in recent years has been discussed. In addition to that it's been discussed how this model can be implemented. At the end, evaluation of the performance of the model using standard evaluation matrices is done.

[10] Deep Reinforcement Learning-Based Image Captioning with Embedding Reward

Image captioning is a challenging problem owing to the complexity in understanding the image content and diverse ways of describing it in natural language. Recent advances in deep neural networks have substantially improved the performance of this task. Most state of-the-art approaches follow an encoder-decoder framework, which generates captions using a sequential recurrent prediction model. However, in this paper, a unique decision making framework for image captioning is introduced by providing the confidence of predicting the next word according to the 18 current states. Additionally, the value network serves as a global and lookahead guidance by evaluating all possible extensions of the current state. In essence, it adjusts the goal of predicting the correct words towards the goal of generating captions similar to the ground truth captions. Both networks are trained using an actor-critic reinforcement learning model, with a novel reward defined by visual-semantic embedding. Extensive experiments and analyses on the Microsoft COCO dataset show that the proposed framework outperforms state-induced. A policy network and a value network is utilized to collaboratively generate captions. The policy network serves as a local gf-the-art approach across different evaluation metrics.

III. GAPS AND ISSUES

The methods have three main issues:

(i) They are trained using maximum likelihood estimation and back- propagation approaches. In this case, the next word is predicted given the image and all the previously generated ground truth words. Therefore, the generated captions look-like ground-truth captions. This phenomenon is called the exposure bias problem.

(ii) Evaluation metrics at test time are non-differentiable. Ideally sequence models for image captioning should be trained to avoid exposure bias and directly optimize metrics for the test time. In the actor-critic-based reinforcement learning algorithm, critique can be used in estimating the expected future reward to train the actor.

(iii) A CNN and RNN based combined network generates captions. We have also included LSTM to enhance the prediction accuracy of our model. LSTMs provide us with a large range of parameters such as learning rates, and input and output biases. Hence, no need for fine adjustments. The complexity to update each weight is reduced to O(1) with LSTMs, similar to that of Back Propagation Through Time (BPTT), which is an advantage of our proposed method.

IV. PROPOSED WORK

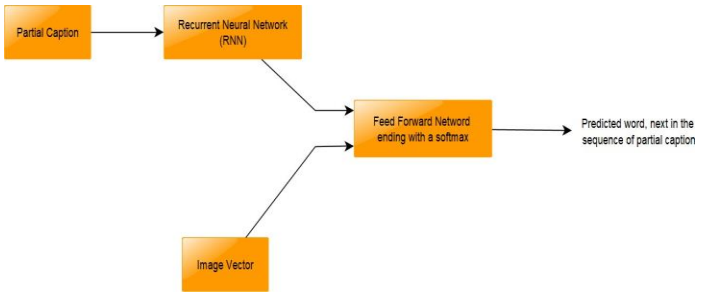


Fig 4.1: Showing the brief architecture which contains the high level sub-modules

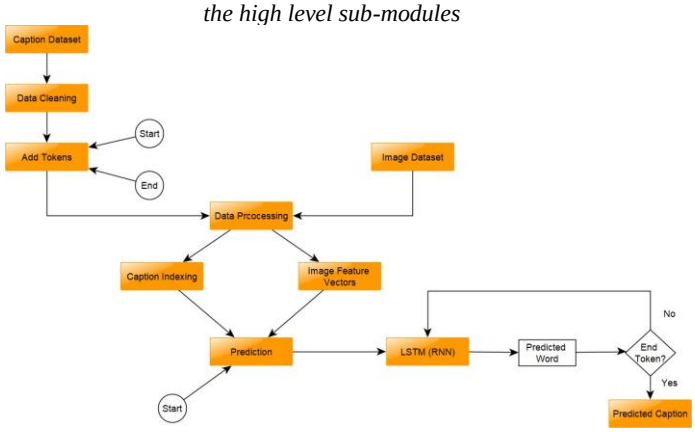


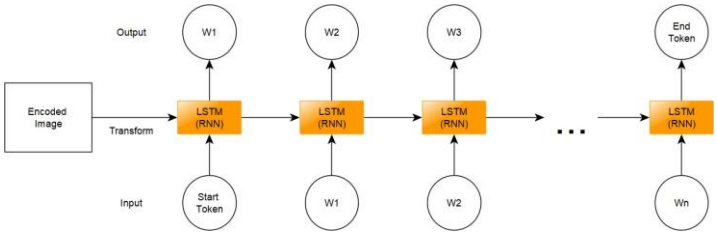
Fig 4.2: Showing the detailed architecture of our proposed model

[1] ENCODER

The Encoder encodes the input image with 3 color channels into a smaller image with "learned" channels. This smaller encoded image is a summary representation of all that's useful in the original image. ResNet 50 is used to encode the images.

[2] DECODER

The Decoder's job is to look at the encoded image and generate a caption word by word. Since it's generating a sequence, it would need to be a Recurrent Neural Network (RNN). LSTM is used for the same.



International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 11 Issue 02, February-2022

[3] Methodology

Image captioning can be divided into three components. The first component includes cleaning the data which consists of captions from the training dataset. Then Start and End tokens are added to the start and end of each caption. The second component includes data processing. Encoding of captions is done by indexing every word to a number and vice-versa. Images are converted to their corresponding feature vectors. Generator function is used to train and test the model wherein the two images and their captions will be used to train and one image to test the model. The third component includes the Decoder which is used in prediction. For prediction, an image vector is provided along with the start token. The model then predicts the next word using LSTM. Then the model uses the predicted words to predict further. This process continues until the end token is predicted.

V. IMPLEMENTATION

There are many steps that need to be done before the model is trained to predict the captions based on the images.

5.1 Data Collection:

There are many open source datasets available like Flickr 8k (containing 8k images), Flickr 30k (containing 30k images), MS COCO (containing 180k images), etc. The Flickr 8k dataset is used in this project. This dataset contains 8000 images each with 5 captions.

5.2. Understanding the data:

"Flickr8k.token.txt" which contains the name of each image along with its 5 captions. A dictionary is created which contains the name of the image (without the .jpg extension) as keys and a list of the 5 captions for the corresponding image as values.

5.3.Data Cleaning:

When dealing with text, generally some basic cleaning is performed like lower- casing all the words ,removing special tokens , eliminating words which contain numbers . All of this would be done by using Natural Language Processing(NLP). There is a special library used for this purpose called nltk.

5.4.Loading the training set:

The text file "Flickr_8k.trainImages.txt" contains the names of the images that belong to the training set. So these names are loaded into a list "train". Now, the descriptions of these images are loaded in the Python dictionary "train_descriptions". However, when they are loaded, two tokens will be added in every caption as follows: 'startseq' -> This is a start sequence token which will be added at the start of every caption. 'endseq' -> This is an end sequence token which will be added at the end of every caption.

5.5.Data Preprocessing — Images:

Images are input (X) to the model. Any input to a model must be given in the form of a vector. There is a need to convert every image into a fixed sized vector which can then be fed as input to the neural network. For this purpose, transfer learning is opted by using the ResNet model (Convolutional Neural Network) created by Google Research. When only the parameters in the last few layers are trained, while the parameters of the earlier layers are only fine-tuned to better fit the purpose. Thus, this training occurs faster. This technique is also referred to as Transfer Learning. This model was trained on Image net dataset to perform image classification on 1000 different classes of images. However, the purpose here is not to classify the image but just get fixed-length informative vector for each image. This process is called automatic feature engineering. All the bottleneck train features are saved in a Python dictionary whose keys are image names and values are corresponding 50 length feature vector. Similarly all the test images are then encoded.

5.6.Data Preprocessing — Captions:

The prediction of the entire caption, given the image does not happen at once. The prediction of the caption word by word. Thus, there is the need to encode each word into a fixed sized vector. Before this, every unique word in the vocabulary will be represented by an integer (index). After this , the maximum length of a caption will be calculated from all the lengths.

5.7.Final Data Preparation for Training and Predicting:

Data will be given as input to the deep learning model that will be used to train and predict the captions corresponding to the pictures or images provided. First, the images would be needed to be converted to their corresponding 50 length feature vector as mentioned above. Secondly, the vocabulary for the train captions would be built by adding the two tokens "startseq" and "endseq" in both of them. Assuming the basic cleaning steps using Natural Language Processing via their libraries has already been performed. Then an index would be given to each word in the vocabulary. After that , there is the need to frame it as a supervised learning problem where the a set of data points D = {Xi, Yi} is available, where Xi is the feature vector of data point 'i' and Yi is the corresponding target variable. Image vector is the input and the caption is what needs to be predicted. But the way the caption is predicted is by predicting the following words based on the current word and the sequence. For the first time, the image vector is provided and the first word as input and tries to predict the second word. One image+caption is not a single data point but multiple data points depending on the length of the caption. So it's typically the partial caption that is going to be generated step by step. In every data point, it's not just the image which goes as input to the system, but also, a partial caption which helps to predict the next word in the sequence. Since sequences are being processed, a Recurrent Neural Network will be employed to read these partial captions. The actual English text of the caption is not going to be passed, rather the sequence of indices where each index represents a unique word is going to be passed. As an index for each word has been created, the words will be replaced with their indices and how the data matrix will look like will be understood. Batch processing would be done in this process ,hence there is a need to make sure that each sequence is of equal length. Hence 0's (zero padding) would be needed to be

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 11 Issue 02, February-2022

appended at the end of each sequence. For this , the maximum length of a caption is calculated. So those many numbers of zeros will be append which will lead to every sequence having the maximum length. Every word (or index) will be mapped (embedded) to higher dimensional space through one of the word embedding techniques. During the model building stage, each word/index is mapped to a 50-long vector using a pre-trained GLOVE word embedding model. In this process ,There is a pretty huge requirement to manage to load this much data into the RAM, it will make the system very slow. For this reason data generators are used a lot in Deep Learning. Data Generators are a functionality which is natively implemented in Python. The ImageDataGenerator class provided by the Keras API is an implementation of the generator function in Python. With the help of this, there is no requirement to store the entire dataset in the memory at once. Even if the current batch of points is available in the memory, it is sufficient for the purpose. A generator function in Python is used exactly for this purpose. It's like an iterator which resumes the functionality from the point it left the last time it was called. After this, the input and the trained vectors will be used to predict the captions of the corresponding images.

5.8.Hyper parameters during training

The model was then trained for 20 epochs with the initial learning rate of 0.001 and 3 pictures per batch (batch size). However after 20 epochs, the learning rate was reduced to 0.0001 and the model was trained on 6 pictures per batch. This generally makes sense because during the later stages of training, since the model is moving towards convergence, we must lower the learning rate so that we take smaller steps towards the minima. Also increasing the batch size over time helps your gradient updates to be more powerful.

VI.RESULTS AND DISCUSSION

As we generate a caption, word by word, you can see the model's gaze shifting across the image. This is possible because of its Attention mechanism, which allows it to focus on the part of the image most relevant to the word it is going to utter next. Here are some captions generated on test images not seen during training or validation. These images are a part of the Flickr8K dataset. To understand how good the model is, let's try to generate captions on images from the test dataset (i.e. the images which the model did not see during the training).

Figure 1:- Output 1

Figure 2 :- Output 2

VII. NOVELTY

We present a model that generates natural language descriptions of images and their regions. Our approach leverages datasets of images and their sentence descriptions to learn about the inter-modal correspondences between language and visual data. Our alignment model is based on a novel combination of Convolutional Neural Networks over image regions, bidirectional Recurrent Neural Networks over sentences, and a structured objective that aligns the two modalities through a multimodal embedding. We then describe a Multimodal Recurrent Neural Network architecture that uses the inferred alignments to learn to generate novel descriptions of image regions. Using one of

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 11 Issue 02, February-2022

the pre-trained models from Keras application APIs, We have done Transfer Learning which gives quite impressive results. Machine learning's potential is often stymied by its heavy reliance on large amounts of high-quality data (for supervised learning this data must be well-labeled to boot). Unfortunately, these sorts of data sets are increasingly proprietary or prohibitively expensive to access— and that's when the necessary data exists at all. Transfer learning allows developers to circumvent the need for lots of new data. A model that has already been trained on a task for which labeled training data is plentiful will be able to handle a new but similar task with far less data. Using a pre- trained model often speeds up the process of training the model on a new task, and can also result in a more accurate and effective model overall. When we augment the knowledge of an isolated learner (also known as an ignorant learner) with knowledge from a source model, the baseline performance might improve due to this knowledge transfer. Utilizing knowledge from a source model might also help in fully learning the target task, as compared to a target model that learns from scratch. This, in turn, results in improvements in the overall time taken to develop/learn a model.

VIII. CONCLUSION AND FUTURE WORK

Automatic image captioning is a relatively new task, thanks to the efforts made by researchers in this field, great progress has been made. In our opinion there is still much room to improve the performance of image captioning. First, with the fast development of deep neural networks, employing more powerful network structures as language models and/or visual models will undoubtedly improve the performance of image description generation. Second, because images are consisted of objects distributed in space, while image captions are sequences of words, investigation on presence and order of visual concepts in image captions are important for image captioning. Furthermore, since this problem fits well with the attention mechanism and attention mechanism is suggested to run the range of AI related tasks, how to utilize attention mechanism to generate image cations effectively will continue to be an important research topic. Third, due to the lack of paired image-sentence training set, research on utilizing unsupervised data, either from images alone or text alone, to improve image captioning will be promising. Fourth, current approaches mainly focus on generating captions that are general about image contents. However, to describe images at a human level and to be applicable in real-life environments, image description should be well grounded by the elements of the images. Therefore, image captioning grounded by image regions will be one of the future research directions. Fifth, so far, most of the previous methods are designed to image captioning for generic cases, while task-specific image captioning is needed in certain cases. Research on solving image captioning problems in various special cases will also be interesting. As a future scope, using a larger dataset would be a better option. Changing the model architecture, e.g. include an attention module could be a better solution for image captioning process. Doing more hyper parameter tuning (learning rate, batch size, number of layers, number of units, dropout rate, batch normalization etc.) and using the cross validation set to understand overfitting can be beneficial. By using Beam Search instead of Greedy Search during Inference and by using BLEU Score to evaluate and measure the performance of the model could give more accurate results.

ACKNOWLEDGMENT

The authors are thankful to VIT, Vellore for giving opportunity to write about such a great and interesting topic.

REFERENCES

[1] Shi, H., Li, P., Wang, B., & Wang, Z. (2018, August). Image captioning based on deep reinforcement learning. In Proceedings of the 10th International Conference on Internet Multimedia Computing and Service (pp. 1-5).
[2] Fu, K., Li, J., Jin, J., & Zhang, C. (2018). Image-text surgery: Efficient concept learning in image captioning by generating pseudopairs. IEEE transactions on neural networks and learning systems, 29(12), 5910-5921.
[3] Yao, T., Pan, Y., Li, Y., Qiu, Z., & Mei, T. (2017). Boosting image captioning with attributes. In Proceedings of the IEEE International Conference on Computer Vision (pp. 4894-4902).
[4] Gao, L., Wang, B., & Wang, W. (2018, February). Image captioning with scene-graph based semantic concepts. In Proceedings of the 2018 10th International Conference on Machine Learning and Computing (pp. 225-229).
[5] Fang, F., Wang, H., & Tang, P. (2018, October). Image captioning with word level attention. In 2018 25th IEEE International Conference on Image Processing (ICIP) (pp. 1278-1282). IEEE.
[6] Wang, M., Song, L., Yang, X., & Luo, C. (2016, September). A parallel-fusion RNN-LSTM architecture for image caption generation. In 2016 IEEE International Conference on Image Processing (ICIP) (pp. 4448-4452). IEEE.
[7] He, C., & Hu, H. (2019). Image captioning with visual-semantic double attention. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), 15(1), 1- 16.
[8] Wang, A., Hu, H., & Yang, L. (2018). Image captioning with Affective guiding and selective attention. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), 14(3), 1-15.
[9] Shah, P., Bakrola, V., & Pati, S. (2017, March). Image captioning using deep neural architectures. In 2017 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS) (pp. 1-4). IEEE.
[10] Ren, Z., Wang, X., Zhang, N., Lv, X., & Li, L. J. (2017). Deep reinforcement learning-based image captioning with embedding reward. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 290-298).
[11] 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).
[12] 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.