International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING

ISSN:2147-6799 www.ijisae.org Original Research Paper

Image Caption Generation Using Recurrent Convolutional Neural Network

[1]BV Subba Rao, [2]K. Meenakshi, [3]K. Kalaiarasi, [4]Ramesh Babu P., [5]J. Kavitha, [6]V. Saravanan

Submitted: 01/10/2023 **Revised**: 21/11/2023 **Accepted**: 01/12/2023

Abstract: This paper presents a residual learning (RL) approach to generate automated captions for any given image. In this approach, a convolutional neural network (CNN) is employed to extract the spectral and spatial characteristics of the image, which is essential to solve the caption generation problem, which necessitates the use of CNN. In addition to this, we consider the nuanced quality of language by incorporating an image annotation generator into the system that has been recommended. The results of the experiments that have been presented here provide convincing evidence that the developed model is an improvement upon the various approaches to image captioning that are currently being used.

Keywords: *Image Captioning, Recurrent neural network, convolutional layers*

1. Introduction

it is not uncommon for a single image to contain a sizeable quantity of data within its confines. On a consistent basis, enormous quantities of visual data are produced by both social media platforms and astronomical observatories. Deep learning gives us the ability to label images in a manner that will eventually eliminate the need for human assistance in this process [1].

There will no longer be a need for human annotations. This will result in a significant reduction in the amount of labor that is required as a direct result of the fact that there will no longer be a need for human participation. The ever-increasing amount of visual material and the accompanying writing that can be found on the internet is one source of the difficulty that users must contend with.

Another source of difficulty is the ever-changing nature of the visual material. Most of these records, unfortunately, have far too much noise to be used in their original state in a image captioning model. This model requires records to be free of noise to function properly [2].

For the purposes of training, models that generate captions require access to a large dataset that is populated with images that have been accurately annotated. This article goal is to demonstrate a system that can produce contextual descriptions of objects discovered in images in an automated manner. These descriptions will be the focus of the article. Consider the information that is shown in the image, and then write a sentence that describes what is taking place there.

This is because the model needs to first extract features from the images, and then use those features to construct a meaningful sentence. As a result of this, this issue has arisen. As a direct consequence of this, the method of automatically generating significant captions is extremely challenging. We can accomplish feature extraction by first introducing a massive dataset into a convolutional neural network (CNN) training environment and then performing many iterations in both the forward and backward directions to identify the optimal weights for the network [3].

This allows us to determine which features should be extracted from the data. After that, a sentence is constructed by utilizing the recurrent neural network (RNN) in addition to the characteristics that were retrieved [4].

We present a residual learning approach to generate automated captions for any given image. The algorithm uses a CNN to extract the spectral and geographic characteristics of an image. In addition, the algorithm can generate a English description of the image with no assistance from a person at all. This is achieved by

[1]*Professor, Department of Information Technology, PVP Siddhartha Institute of Technology, Vijayawada, Andhra Pradesh, India. Email ID: bvsrau@gmail.com*

[2]*Professor, Department of Mathematics, VTU(RC), CMR Institute of Technology, Bengaluru, Karnataka, India. Email ID: meenakshik531@gmail.com*

[3]*Assistant Professor, PG and Research Department of Mathematics, Cauvery College for Women (Autonomous), Tiruchirappalli, Tamil Nadu, India. Email ID: kalaishruthi12@gmail.com*

[4]*Associate Professor, Department of Computer Science, College of Engineering and Technology, Wollega University, Nekemte, Oromia Region, Ethiopia. Email ID: drprb2009@gmail.com*

[5]*Associate Professor, Department of Basic Sciences, Cambridge Institute of Technology (CIT), Bengaluru, India. Email ID: kavitha.maths@cambridge.edu.in*

[6]*Associate Professor, Department of Computer Science, College of Engineering and Technology, Dambi Dollo University, Dambi Dollo, Oromia Region, Ethiopia. Email ID: reachvsaravanan@gmail.com*

combining two methods, namely, the encoder-decoder strategy and the focused concentration technique.

2. Related works

The RNN in most of the models that are presently available by employing Maximum Likelihood Estimation (MLE) to produce image descriptions. This was done to improve accuracy. The reasoning that is presented in [8] suggests that MLE methods are susceptible to exposure bias while the inference phase is being carried out.

When it comes to the captioning of images, the MLE runs into the same issue that it did in the past, which is that the outcomes do not agree with human judgments of quality. Within the confines of the generative adversarial network (CNN) paradigm, the MLE can be changed out for a different strategy that is accessible [9]. The MLE can be replaced with this different design as an alternative. CNN was originally conceived with the intention of producing fake images that were convincing to the human eye [5].

Generative adversarial networks, also known as CNNs, can learn generative models without the need for the loss function of the target distribution to be specified in preparation. This enables CNNs to learn more quickly than traditional machine learning methods. Instead, a discriminator network that looks for differences between actual and generated samples is a part of CNN. This is done to achieve higher levels of precision. When a network is being trained, a technique known as adversarial training is applied to the entire system so that it can learn [6].

The individual can then construct a discriminator to evaluate the genuineness of the instances that were generated by the caption generator. This can be done in several ways. The generator that is used by CNN, which generates a caption based on the characteristics of the image, is theoretically comparable to the generator that is used by the caption. When attempting to find a solution to a problem that involves language, CNN has a error that needs to be worked around before it can be used effectively [7].

Language difficulties are fundamentally distinct from visual disabilities in several important ways. Because there are so few of these tokens, the gradients cannot backspread through them even if they are passed directly into the discriminator. This is because the discriminator is a binary machine. One approach that may be put into action to ascertain the gradients of the discontinuous units is to make use of a reinforcement learning (RL) framework [8].

According to [9], the RL framework has a problem because there is no intermediate reward when it comes to the generation of sequences. This problem is described in more detail in the previous paragraph. You will not be provided any indication regarding the reward until all the tasks that need to be completed by you have been successfully completed. For us to be able to optimize the complete sequence, we need to know the long-term benefit that will be conferred by each token that is generated at an intermediate stage.

We have developed a framework for image captioning that is based on the generalized additive model (CNN) to address the problems that were brought up previously in the conversation. The discriminator that is used in the system that has been recommended considers the degree to which freshly generated captions resemble the reference captions as well as the degree to which they are consistent with the image features. Both factors are taken into consideration.

The networks can adjust take into consideration the possibility that it will generate subtitles that are unrealistic if they conduct an analysis of the discriminator. In addition to this, we take into consideration the nuanced quality of language by incorporating an image annotation generator into our RL system. This allows us to reflect the linguistic landscape more accurately. A generator is rewarded for their efforts by a discriminator through the statements that are made by the discriminator. By utilizing a Bayesian network, this technique allows us to modify the parameters of the image captioning generator. A stochastic parameterized policy serves as the inspiration for the design of the image captioning generator.

The research makes use of Policy Gradient, which swiftly addresses the differential problems that are present in conventional CNN, to train the policy network. It is another source from which we have taken inspiration for this concept. We put this plan into action so that we can address the shortage of intermediary rewards that has been causing us problems.

This is utilized by the software to get a taste of the potential long-term benefit of an intermediary step. If we consider the generation of sequence tokens to be the task that needs to be completed, then we can acquire the intermediary rewards in RL by employing a strategy that is very similar to the Monte Carlo roll-out approach. In this work, we make use of a similar sampling approach to manage the intermediate benefits that come along with the process of caption generation. This is done to ensure the integrity of the final product.

3. Feature Map Generation using CNN-GAN

There are many levels of complexity that need to be taken into consideration whenever a computer attempts to interpret an image. The data from the images that are taken in by an animal iris are then sent to the brain, where the neurons conduct analysis on the information that has been received. It is motivated by the design of the animal visual cortex and aims to learn spatial hierarchies of characteristics, beginning autonomously and adaptively with the most basic and progressing to the most complex.

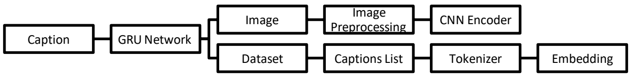

Fig 1: Proposed Image Captioning

The architectural structure of a typical CNN consists of several different levels, including convolution, pooling, and completely connected levels. The first two layers, which are known as convolution and pooling, are the ones that oversee the process of extracting features from the data. These features are then passed on to the third layer, which is a completely connected one, to be used in the output of the model. Convolution is one of the most important mathematical operations that goes into the making of a CNN, which is constructed using a wide variety of other mathematical operations. The term convolution refers to a specific type of linear operation.

A 3D kernel is applied to the images to extract the spectral and spatial characteristics, which is essential to solve the HSI classification problem, which necessitates the use of CNN. This step is necessary to accomplish the goal of applying the CNN to the problem. To performing integrated feature mapping on raw data that possesses both spectral and geographic dimensions, it is feasible to make use of three-dimensional convolutional kernels. This can be done. The formula for 3D convolution is:

$$v_{ij}^{xy} = \Phi\left(b_{ij} + \sum_{m} \sum_{p=0}^{P_i-1} \sum_{q=0}^{Q_i-1} w_{ijm}^{pq} \times v_{(i-1)m}^{(x+p)(y+q)} \right)$$

where the value under the current operation is at the position (*x*, *y*, *z*),

i - current layer,

j - feature map,

v_{ij}^{xy} - output,

r - bias term,

k_{ijp}^{hwb} - weight value,

p - features set,

f - activation function,

The CNNs such as 3D CNN is used to evaluate the performance of each dimensionality reduction technique on three publicly available hyperspectral datasets and are discussed in the following sections.

4. Caption generation using Residual Learning

Consider H(x) as a mapping that a few stacked layers, but not necessarily the entire net, can fit into, with x representing the inputs to the first of these layers. This is just one way to think about H(x). This is merely one perspective among many on the matter. If a group of nonlinear layers can asymptotically approximate a set of complex functions2 is the same as assuming that those layers can asymptotically approximate the residual functions, H(x) x. This is because both assumptions are equivalent. These two possibilities are completely interchangeable with one another.

This is the reason why we don't expect stacked layers to be able to approximate H; instead, we enable them to approximate a residual function called F(x):= H(x) x. This is because we don't expect stacked layers to be able to approximate H. (x). After making these modifications to the fundamental function, the structure that is produced is denoted by the symbol F(x+x). It is possible that the relative ease with which each representation can be acquired will be different, even though it is anticipated that both representations will be able to approximate the objective functions.

The intrinsic ambiguity that can be found within the degradation problem served as the driving force behind this clarification. If the additional layers can be constructed as identity mappings, as we mentioned in the introduction, then the training error for a deeper model shouldn't be any higher than that of its shallower equivalent. This is because deeper models require more data to train than shallower models. If the underlying model can be constructed, this should be the expected outcome.

Experiments have shown that identity mappings are a suitable technique of preconditioning since the learned residual functions have relatively modest responses on average. This is because identity mappings map the optimal function to a point that is closer to an identity mapping than it is to a zero mapping.

3.2.1. Identity Mapping

Between each successive layer of layering, we use residual learning at predetermined intervals.

In this research, we are going to start with the officially specified building block y = F(x, Wi) + x as our point of departure. This will be the case throughout the entirety of this investigation.

$$y = F(x, \{Wi\}) + x.$$

In this scenario, the input and output vectors of the levels that are being considered are represented by the x and y, respectively. The part of the mapping that has not yet been understood is represented by F, which stands for the function. (x, Wi).

F = W2(W1x),

where RLU is depicted, and the biases have been omitted for the purpose of keeping things as simple as possible. Finding the expression F := x requires performing both a quick link and an addition operation on each individual constituent to arrive at the correct answer. The second type of nonlinearity, which is represented by the letter y, is the one that we focus on.

The simplifications made to the connections in Equation (1) do not result in the addition of any new parameters and do not in any way make the calculations more difficult. This is beneficial not only when used in day-to-day situations but also when comparing standard networks to leftover networks. This is important when comparing standard networks to leftover networks. If the same number of parameters, depth, width, and computational expense are used in each instance, then it is possible to make valid comparisons between plain networks and residual networks.

It is important that the values x and F in the equation have the same quantity of weight and significance. If this is not the case, however (for example, when transitioning between input and output channels), it is still possible to carry out a linear projection Ws by making use of the short-cut connections, as follows:

$$y = F(x, \{Wi\}) + Wsx$$

There is also the possibility of utilizing a square collection Ws in the solution that was presented earlier. The research demonstrate through experimentation that identity mapping is sufficient for addressing the deterioration problem and that it is cost-effective; as a result, Ws will only be used when matching dimensions.

There is a wide range of possibilities for the shape that the residual function F can assume. Even though it is theoretically possible for function F to have a greater number of layers than the two or three that are depicted in figure 5, we conduct all our experiments with a function F that has either two or three layers throughout the duration of this article. If, on the other hand, F consists of only a single layer, then equation (1) is quite comparable to the equation that describes a linear layer, which is written as follows: y is equal to W1x plus x, but we have not found any benefits associated with using this kind of layer.

The notations that were discussed earlier are applicable to convolutional layers, which is an important point that we want to bring to your attention. Even though they were originally developed for use with fully connected layers, you can use them with convolutional layers. A substantial number of a convolutional neural network layers can be represented by the function F. The components of two feature maps are merged together, channel by channel, after being added to one another element by element.

5. Results and Discussions

To verify the validity of the experimental proof, we will be utilizing the MS-COCO 2014 data accumulation in conjunction with the Flick8K dataset. In addition to that, comparisons are made between older versions and more current ones. There are a total of 8,000 photographs, and each one includes a variety of comments that provide additional insight into the subject matter depicted in the image.

Even when these older, more established methods are used, the CNN architecture still manages to enhance the image captioning process so that it is more accurate. However, the evaluation metrics place an excessive amount of emphasis on n-gram matching and pattern matching with ground truth captions and completely disregard the naturalness of the language, even though these can represent the accuracy of the description related to the image.

In addition, the evaluation metrics place an excessive amount of emphasis on n-gram matching and pattern matching with ground truth captions. Despite the fact that both of these kinds of matching can be used to identify n-grams that are similar to one another, this is the result.

Table 1: Training Results

Methods	METEOR	CIDEr	Rouge-L
KNN	25	123	57
GAN	28	128	66
CNN	24	125	57
RNN	27	124	65
ResNet50	27	137	64
RL	38	141	68
Proposed	42	145	71

Table 2: Testing Results

Methods	METEOR	CIDEr	Rouge-L
KNN	23	181	55
GAN	26	173	47
CNN	21	172	48
RNN	24	180	54
ResNet50	28	195	48
RL	32	197	52
Proposed	34	201	60

Table 1: Training Results of BLEU

Methods	BLEU-1	BLEU-2	BLEU-3	BLEU-4
KNN	56	32	19	12
GAN	76	52	38	28
CNN	81	55	39	27
RNN	72	51	37	26
ResNet50	83	64	47	37
RL	79	60	46	34
Proposed	84	67	49	38

Table 2: Testing Results of BLEU

Methods	BLEU-1	BLEU-2	BLEU-3	BLEU-4
KNN	67	33	27	16
GAN	72	48	31	18
CNN	74	49	32	21
RNN	66	43	29	19
ResNet50	71	52	42	29
RL	76	50	38	26
Proposed	78	53	44	31

The results of the experiments make it feasible to reach the conclusion that the CNN-ResL strategy achieves the MAPE and MAE values that are as low as they can get while still being effective. The findings and the discussion that have been presented here provide convincing evidence that the developed model is an improvement upon the various approaches to image captioning that are currently being used.

CNN method routinely produces the finest results possible across the board because of how it implemented. Nobody ought to be caught off guard by this information. Their usefulness is severely limited since metrics place a greater emphasis on n-gram matching in connection to the references rather than taking into consideration more holistic qualities such as naturalness and diversity. These discoveries have significant ramifications due to the reality that image captions ought to be judged by human beings. In addition, they give the impression that the strategies for image captioning will improve if there is a greater reliance placed on natural expression and variation.

6. Conclusions

We have shown that an image can be analyzed by a CNN algorithm, and that the algorithm can then generate a passable English description of the image with no assistance from a person at all. They begin by compressing an image into a representation with a tool known as a convolutional neural network, or CNN, and then move on to another tool known as a recurrent neural network, or RNN, to construct a statement that is consistent with the image. The algorithm has been fine-tuned so that it can determine the assertion that is most likely to be correct given the image that is presented. In addition to this, we investigated the results of combining two methods, namely the encoder-decoder strategy and the focused concentration technique.

References

[1] Ding, S., Qu, S., Xi, Y., & Wan, S. (2020). Stimulus-driven and concept-driven analysis for image caption generation. *Neurocomputing*, *398*, 520-530.

[2] Chen, S., Jin, Q., Wang, P., & Wu, Q. (2020). Say as you wish: Fine-grained control of image caption generation with abstract scene graphs. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 9962-9971).

[3] Liu, M., Li, L., Hu, H., Guan, W., & Tian, J. (2020). Image caption generation with dual attention mechanism. *Information Processing & Management*, *57*(2), 102178.

[4] He, X., Shi, B., Bai, X., Xia, G. S., Zhang, Z., & Dong, W. (2019). Image caption generation with part of speech guidance. *Pattern Recognition Letters*, *119*, 229-237.

[5] Zhou, Z., Zhang, X., Li, Z., Huang, F., & Xu, J. (2022). Multilevel attention networks and policy reinforcement learning for image caption generation. *Big Data*, *10*(6), 481-492.

[6] Agrawal, V., Dhekane, S., Tuniya, N., & Vyas, V. (2021, July). Image Caption Generator Using Attention Mechanism. In *2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT)* (pp. 1-6). IEEE.

[7] Zhao, S., Li, L., Peng, H., Yang, Z., & Zhang, J. (2020). Image caption generation via unified retrieval and generation-based method. *Applied Sciences*, *10*(18), 6235.

[8] Liu, X., & Xu, Q. (2020). Adaptive attention-based high-level semantic introduction for image caption. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, *16*(4), 1-22.

[9] Mounika, S., & Vijaybabu, P. (2022). Image caption generator using cnn and lstm. *South Asian Journal of Engineering and Technology*, *12*(3), 78-86.