Journal of Research and Applications (JRA), UCSMTLA, 2019 Volume-01, Issue-01

Article ID - JRA190042

ANALYSIS ON RESIDUAL NETWORK MODELS FOR IMAGE DESCRIPTION GENERATION

Phyu Phyu Khaing [(1)], **Myint San**[(2)], **Mie Mie Aung**[(2)]

[(1)]Myanmar Institute of Information Technology, Mandalay, Myanmar

[(2)]University of Computer Studies (Monywa), Myanmar

[(1)]phyu_phyu_khaing@miit.edu.mm

ABSTRACT

Image description generation is the process in which the sentence is generated for the context of the image. In this process, there are two important parts: feature extraction from the image and sentence generation based on the image features. The process of image description generation works as the combination of computer vision and natural language processing. This paper intends to represent the image description generation framework based on the pre-trained Convolutional Neural Network trained by Deep Residual Network (ResNet) for feature extraction and Two Long-Short Term Model (Two-LSTM) for sentence description. The experiment of the system is implemented on the Flickr8Kdatasets and the performance of the system is measured using standard evaluation metrics, such as: BLEU, METEOR, CIDEr, and ROUGE-L.

KEYWORDS: *Image Description Generation, Convolutional Neural Network, ResNet, Long-Short Term Memory, LSTM.*

1. INTRODUCTION

Image description generation process is still challenging in the artificial intelligence processes and is used both computer vision process and natural language process. Computer vision is utilized to extract the features from the image, and natural language processing helps in generating the sentence description that further describes what the image shows. The process of the image description generation requires the model to acknowledge objects in an image, to comprehend the relationship between objects and to communicate them in a natural language sentence.

Deep learning model has become as core component any advanced applications, and the usage of this learning method is also much more available. Image captioning can produce a single significant and comprehensive grammar phrase and increase description precision by using a profound learning model. This paper presents the image description generation framework by using deep Residual Network (ResNet) and Long-Short Term Model (LSTM). The system is implemented on the benchmark datasets: Flickr8k; and evaluated the performance of the model using BLEU, METEOR, CIDEr, and ROUGE-L.

The rest of the paper is prepared with the following sections. The related works of this research is reviewed in Section 2. Section 3 shows the structure of the image description generation framework and explains the methodologies. The experimentation is implemented in Section 4. Section 5 concludes about the methods with the limitation and future works of the system.

2. RELATED WORKS

Vinyals et al. [1] suggested an end-to-end neural image caption generator model called NIC by integrating deep CNN with LSTM. NIC first utilizes a convolutional neural network to encode the image and then utilizes a recurrent neural network to produce a natural language sentence that matches with the image. This measured efficiency with the conventional assessment metrics and also assessed individual judgments on five benchmark datasets.

Mao et al. [2] initiated the multimodal recurrent neural network (m-RNN) technique to extract new image caption. The proposed m-RNN technique consists of two networks. A convoluted neural network that encodes images and adeep recurrent neural network generates sentences. The two subnets communicate with each other in the multimodal layer to form the entire m-RNN model. The input to this method is the image and the sentence. The probability distribution is calculated to form the next captioning word for the sentence. The model includes five additional layers: a two-word embedded layer, a recurrent layer, a multimodal layer, and a SoftMax layer.

Donahue et al. [3] proposed the Long-term Recurrent Convolutional Network (LRCN) that is similar with m-RNN [4]. for visual recognition and description, LRCN incorporates long-range temporal recursion with convolutional layers. LRCN uses AlexNet for feature extraction of the image and a stacked two-layer LSTM for image description generation by feeding the image content. In a stacked two-layer LSTM, bottom LSTM is supplied only with the prior word to embed, and the top LSTM takes the feature of the image as input and the bottom LSTM output the result. The benefit of the stacked architecture is that top LSTM merges the image and the context information to predict the distribution of word, and bottom LSTM can concentrate on text information modelling.

Jia et al. [5] suggested the alternative LSTM extension which is called guided LSTM (gLSTM) and long sentences can be generated by using this gLSTM. This model brings the semantic image-extracted information into LSTM cell state and each gate to produce image descriptions as input together with entire image. It also examines the various strategies for controlling the captioning length by normalizing with the semantic information. Multimodal embedding space can also be used to extract semantic based information from the image.

Wang et al. [6] indicated the architecture by integrating the benefits of simple RNN and LSTM used throughout parallel fusion for image captioning, namely parallel-fusion RNN-LSTM. This strategy enhances performance and effectiveness through evaluation on Flickr8 K dataset with BLEU and METEOR. Future research must examine the restriction of simultaneous threads by using more complicated image features to concentrate the greater efficiency.

Aneja et al. [7] studied convolutional image captioning technology, which has proven to be consistent with current LSTM techniques. Validity and comparison performance is shown by verifying on the MSCOCO datasets. Models can be used to carefully improve performance. Wang et al. [8] found a framework for producing the caption of the image using a convolutional neural network (CNN). MSCOCO has been extensively studied to examine the effects of model size and thickness. Compared to the LSTM-based model using a similar attention process, the proposed design can be compared in terms of BLEU, METEOR and CIDER scores.

There are few important literatures on image description processes but mostly lack of explicit concluding remrks and scope for research directions that can be implemented. Thus, an extensive comparison is required to assess what methodology could be considered as benchmark with suitable datasets.

3. IMAGE DESCRIPTION GENERATION MODEL

This system presents the analysis of image description generation model by using deep residual network (ResNet) and long-short term model (LSTM) as shown in Fig 1.

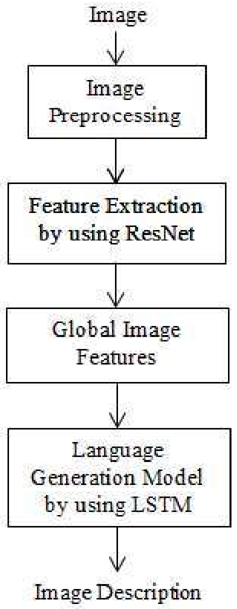


Image
Image Preprocessing
Feature Extraction by using ResNet
Global Image Features
Language Generation Model by using LSTM
Image Description

Fig 1.Image Description Generation Model

3.1 Convolutional Neural Network

Convolutional Neural Network (CNN) is fundamental type of deep learning method similar to ordinary neural network with hidden layers. In convolutional neural network, increasing the number of hidden layers poses several challenges which are overcome in CNN. In CNN, convolutional layers, pooling scheme, normalization scheme and fully connected layers are main components. Pooling is most important component that helps in reducing the repeating number of features and allowing the hidden layers to increase. CNN works much better than ordinary SVM or any other training method.

There are many pre-trained model for Convolutional Neural Network (CNN). Among them, some pre-trained model that contained in Keras. Keras is a neural-network library, which is open-source and is written with Python. There have been many research areas that used Keras. For image captioning, many researchers used Keras pre-trained model for image encoding that extracts the features of image. It supports convolutional neural networks, recurrent neural networks, and also combinations of two networks; and runs impeccably on CPU and GPU. There are many pre-trained models for image captioning. Among them, this study is applied deep Residual Networkfor the feature extraction of the image.

The Residual Network (ResNet) [9] is a specialized neural network that helps to manage more advanced deep learning tasks and models. ResNet is implemented by adding the corresponding residual version to the normal network. In the network, the layer uses the reference as a layer input to learn the redundant function instead of learning the function without reference. ResNet can achieve precision from a greatly increased depth and is easy to optimize. ResNet is available in 18, 34, 50, 101 and 152 versions. 34-layer ResNet is better than 18-layer ResNet. The 34-tier ResNet reduced the previous error by 3.5%. The 18-layer ResNet is faster than an 18-layer normal network. ResNet 50/101/152 is more sensitive than ResNet 34. Therefore, ResNet50, ResNet101, and ResNet152 are implemented in the feature extractor.

3.2Long-Short Term Memory

Long-Short Term Memory (LSTM) is used that can handle long term dependence and also address the problems of hidden states. LSTM was actually suggested by Hachreiter (1997) mainly to deal with long term dependency issues and the information can retain for very long time also [10]. LSTM is designed in chain like structure with four important layers, namely Forget Gate (f); Input Gate (I); Candidate Layer (C); Output Gate (O). Except Candidate layer, other layers are all single layered networks that use Sigmoid function whereas Candidate layer utilizes Tanh function for activation purposes. Generally, three major steps are executed in LSTM operation. The information that has to be thrown away is processed in first step and second step determines what information has to be stored as new information. Output is appropriately decided by third step.The output of the LSTM cell is calculated with the following equations.

$$f_t = \sigma(x_t * W_{xf} + h_{t-1} * W_{xf}) \quad (1)$$

$$\overline{c_t} = \tanh(x_t * W_{xc} + h_{t-1} * W_{hc}) \quad (2)$$

$$i_t = \sigma(x_t * W_{xi} + h_{t-1} * W_{hi}) \quad (3)$$

$$o_t = \sigma(x_t * W_{xo} + h_{t-1} * W_{ho}) \quad (4)$$

$$c_t = f_t * c_{t-1} + i_t * \overline{c_t} \quad (5)$$

$$h_t = o_t * \tanh(c_t) \quad (6)$$

In the equations,f_t is forget gate, $\overline{c}_t$ is cell state, i_t is input gate and o_t is output gate, c_t is current cell memory, h_t is current cell output, c_{t-1} is previous cell memory, h_{t-1} is previous cell output, x_t is input vector. W_x and W_h are the weights of input and the previous cell output respectively.

4. EXPERIMENT

4.1 Dataset

There has been many popular datasets that is used for image description generation model such as Flickr8k [11], Flickr30k [12], MSCOCo [12], etc. Among them, this study presents on Flickr8k [11] benchmark datasets. That is very popular and mostly used for image annotation. The images from Flickr.com website is collected for Flickr8k dataset. Flickr8k dataset consists of 8,092 images; and divided into 6,000 images to train, 1,000 images to test and 1,000 images to validate. Five sentences are created for each image in the dataset by characterizing with events scenes, situation, and entities (animals, people and objects). The grammar of images from the dataset is tested with the help of the workers and spelling is checked with United State format.

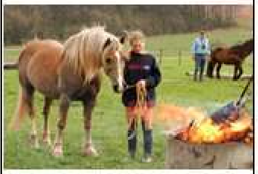
	1. A woman carry a white ball be run behind a small boy . 2. A woman hold a ball chase a small boy run in the grass . 3. A woman hold a small ball chase after a small boy . 4. A woman be run after a boy on the grass . 5. A woman with a softball run after a child in a grassy lawn .
	1. A blond horse and a blond girl in a black sweatshirt be stare at a fire in a barrel . 2. A girl and her horse stand by a fire . 3. A girl hold a horse 's lead behind a fire . 4. A man , and girl and two horse be near a contain fire . 5. Two person and two horse watch a fire .

Fig 2. Sample images and descriptions from the dataset

4.2Evaluation Metrics

The evaluation metrics are mostly implemented to measure the performance of image captioning model. There are many popular evaluation metrics. Among them, this study is evaluated with Bilingual Evaluation Understudy (BLEU), Metric for Evaluation based Image Description Evaluation (METEOR), Consensusbased Image Description Evaluation (CIDEr), and Recall-Oriented Understudy for Gisting Evaluation (ROUGE).

BLEU [13] measures by the similarity between machine-generated sentences and sentences in a data set. It is part of the N-gram. This is very similar to human judgment. Therefore, this is often referred to as a humanoid assessment for machine translation. In this experiment, the performance of the model will be verified at 1, 2, 3 and 4 grams, respectively.

To overcome the BLEU metric problem, METEOR [14] was designed based on the harmonic averages of unigram precision and recall rates. It is used the same resources to evaluate the target language and create a statistical translation system. Open source software is freely available for that. Unlike BLEU indicators, METEOR seeks compliance at the corpus level.

CIDEr [15] is used to evaluate definitions created using human consensus. The purpose of CIDEr is to automatically evaluate the description for the image. The metric specifies the number of units of ideas that match the candidate sentence in the image description set.

ROUGE-L [16] is designed for measurement of a general result of maximum length between the target and source statements. There are four different ROUGE measurements: ROUGE-L, ROUGE-N, ROUGE-S, and ROUGE-W. Because the survey has a sentence-level LCS, ROUGE-L is used for the most common sequence of results (LCS) in this experiment.

4.3 Experimental Results

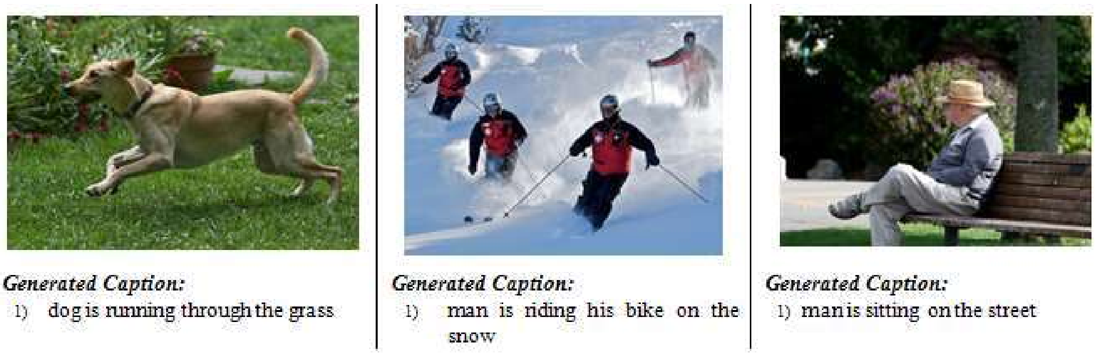
To implement the image description generation model, we have used the machine with Intel Core I7 processor with 8 cores and 8GB RAM running on Window 10 OS. Keras library based on tensorflow is used for creating and training deep neural networks. Tensorflow is a deep learning library developed by Google [17]. Tensorflow uses the graph definition to implement the deep learning network. It can be executed on any supported devices by defining one graph at once.

For the image description generation model, we are using by combining CNN and LSTM. ResNetsare used for the image features extraction task and act as an image encoder. That model is trained with 10 epochs. Table 1 shows the experimental results for image description generation.Generally, ResNet50 is better than the other two and ResNet152 is better than ResNet101. However, BLEU-2, BLEU-3, BLEU-4 of ResNet152 is the better than ResNet50 and ResNet101. Sample generated descriptions of the images are shown in Fig.3.

This work compares few image description generation models in terms of various performance metrics and it can be evidently seen that ResNet50 outperforms all the models. So, the comparison of literature suggests that we need to expensively evaluate and based on that the suitable model can be recommended.

Table 1. Experimental results of image description generation

Models	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	CIDEr	ROUGE_L
ResNet50	**0.576308**	**0.305348**	**0.167662**	**0.085092**	**0.207652**	**0.213671**	**0.454116**
ResNet101	0.568494	0.303183	0.167712	0.088484	0.201189	0.188736	0.447229
ResNet152	0.575645	0.31071	0.169603	0.087169	0.20544	0.197922	0.453826



Generated Caption:
1) dog is running through the grass

Generated Caption:
1) man is riding his bike on the snow

Generated Caption:
1) man is sitting on the street

Fig 3.Some Generated Captions for Sample Images

5. CONCLUSIONS

This paper studies the analysis for image description generation models and the study used the encoder-decoder framework for image description generation. Deep residual networks (ResNets)are used as the encoder to extract features from image and LSTM is used as the decoder to generate a sentence that described the image. This study is implemented on Flickr8k dataset and evaluate with BLEU, METEOR, CIDEr, and ROUGE-L evaluation metrics. We have evaluated different models in term of performance evaluation parameters and observed that the model ResNet50 does well in comparison with other models.The work, however, doesn't consider the attention mechanism on the model. For the extension, we will test with different datasets and other different models, and we will also add the attention models in the encoder-decoder framework. Though, the test and experimentation were carried out for a dataset which is significantly used by numerous researchers.

ACKNOWLEDMENT

The authors are grateful for the constructive advice and comments to all the esteemed reviewers who have anonymously helped a lot for making the manuscript of very good quality.

REFERENCES

[1] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, USA, 2015, pp. 3156–3164.

[2] J. Mao, W. Xu, Y. Yang, J. Wang, Z. Huang, and A. Yuille, "Deep Captioning with Multimodal Recurrent Neural Networks (m-RNN)," *ArXiv14126632 Cs*, Jun. 2015.

[3] J. Donahue *et al.*, "Long-Term Recurrent Convolutional Networks for Visual Recognition and Description," p. 10.

[4] T.-Y. Lin *et al.*, "Microsoft COCO: Common Objects in Context," in *Computer Vision – ECCV 2014*, vol. 8693, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds. Cham: Springer International Publishing, 2014, pp. 740–755.

[5] X. Jia, E. Gavves, B. Fernando, and T. Tuytelaars, "Guiding Long-Short Term Memory for Image

Caption Generation," Sep. 2015.

[6] M. Wang, L. Song, X. Yang, and C. Luo, "A parallel-fusion RNN-LSTM architecture for image caption generation," in *2016 IEEE International Conference on Image Processing (ICIP)*, 2016, pp. 4448–4452.

[7] J. Aneja, A. Deshpande, and A. G. Schwing, "Convolutional Image Captioning," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, 2018, pp. 5561–5570.

[8] Q. Wang and A. B. Chan, "CNN+CNN: Convolutional Decoders for Image Captioning," *ArXiv180509019 Cs*, May 2018.

[9] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, 2016, pp. 770–778.

[10] S. Hochreiter and J. Schmidhuber, "Long Short-term Memory," *Neural Comput.*, vol. 9, pp. 1735–80, Dec. 1997.

[11] M. Hodosh, P. Young, and J. Hockenmaier, "Framing Image Description as a Ranking Task Data, Models and Evaluation Metrics Extended Abstract," p. 5.

[12] B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier, and S. Lazebnik, "Flickr30k Entities: Collecting Region-to-Phrase Correspondences for Richer Image-to-Sentence Models," p. 9.

[13] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: a method for automatic evaluation of machine translation," in *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02*, Philadelphia, Pennsylvania, 2001, p. 311.

[14] M. Denkowski and A. Lavie, "Meteor Universal: Language Specific Translation Evaluation for Any Target Language," in *Proceedings of the Ninth Workshop on Statistical Machine Translation*, Baltimore, Maryland, USA, 2014, pp. 376–380.

[15] R. Vedantam, C. L. Zitnick, and D. Parikh, "CIDEr: Consensus-based Image Description Evaluation," *ArXiv14115726 Cs*, Jun. 2015.

[16] C.-Y. Lin, "ROUGE: A Package for Automatic Evaluation of Summaries," in *Text Summarization Branches Out*, Barcelona, Spain, 2004, pp. 74–81.

[17] M. Abadi *et al.*, "TensorFlow: A system for large-scale machine learning," p. 21.