Grenze International Journal of Engineering and Technology, Jan Issue

CNN-GRU: Transforming Image into Sentence using GRU and Attention Mechanism

Gurdeep Saini[1] and Nagamma Patil[2]
[1-2]National Institute of Technology, Surathkal, Karnataka
Email: gsaini.cse@gmail.com, nagammapatil@nitk.edu.in

***Abstract*—Recent advancement of the deep neural network has triggered great attention in both Natural Language Processing (NLP) and Computer Vision (CV). It provides an efficient way of understanding semantic and syntactic structure which can deal with complex task such as automatic image captioning. Image captioning methodology mainly based on the encoder-decoder approach. In the present work, we developed a CNN-GRU model using Convolutional Neural Network (CNN), Gated Recurrent Unit (GRU) and attention mechanism. Here VGG16 is used as an encoder, GRU and attention mechanism are used as a decoder. Our model has shown significant improvement compared to other state-of-art encoder-decoder models on the famous MSCOCO data set. Further, the time taken to train and test our model is two-third as compared to other similar models such as CNN-CNN and CNN-RNN.**

***Index Terms*— Computer Vision, Image Captioning, Machine Translation, Natural Language Processing, Video Captioning.**

I. INTRODUCTION

Transforming an image into a sentence is a way to generate the sentence for an image. Key points of sentence generation are to identify the objects, actions and hidden features in the images, then stabilize the relationship between them to form a meaningful sentence. After identifying the objects and actions in the image, next step is to generate the syntactically and semantically correct sentence. Image Captioning is combination of following two task.

1. Computer vision for object identification.
2. Natural Language Processing (NLP) for sentence generation.

It is a very challenging task to imitate human brain's ability, but many researchers have shown significant achievement in this field [12, 13]. Deep learning based techniques can handle this task using CNN and RNN. Image captioning can be used in image retrieval using text, many intelligent control systems like locomotives and radar, internet of thing (IoT) based devices [20], caption generation for medical images, information accessed for blind users, self-driving car and human-robot interaction. With the evolution of deep learning model image captioning has evolved a new research field because of its applications.

There are two methods of image captioning one is bottom-up approach [1, 3] and second is a top-down approach [4, 5]. In the Bottom-up approach [1, 3] first identify the object then combine them in a meaningful way, but in case of top-down approach [4, 6] it first made the semantic representation of image then semantic representation is decoded into sentence using various deep neural network techniques such as CNN, RNN and long short term memory (LSTM). Neural network and multi-model techniques are more efficient way to caption generation but we need to improve those techniques for better result.

Grenze ID: 01.GIJET.7.1.8_2

Existing methodologies have less accuracy and consume much time for training and testing. Our work has taken less time and better performance as compared to existing models. It is based on the top-down approach in which first we will use the convolution neural network to generate features of the image then we will decode it using attention mechanism and GRU based RNN network to generate a sentence. The main contribution of the our model is as follow.

1. Initially VGG16 [16] is used to identify the object and action in the image.
2. GRU based recurrent neural network and Bahdanau Attention is used as a decoder.
3. We performed the experiment on MSCOCO [22] data set.

II. Related Work

Existing image and video captaining mechanism can be classified into two different classes.

1. Template Based [18]
2. Encoder-Decoder approaches

The first step of Template-based caption generation is splitting the large sentences into the fragments like subject, verb and object using the rules of respective language's grammar. Each generated fragment is linked to an action or object in the image. Now for the test image, sentence is generated using the pretrained template. Therefore generated sentences are highly dependent on the template structure.

Initially, retrieval based captioning was very famous. In this process, generated caption is fetched from the caption pool for the query image. The generated caption may be composed of more than one existing captions. The first step is to find the related image from training data set which is suggested by some re-searchers. For the query image first, solve the Markov random field and plot the meaning space of the image. With the help of Lin similarity measure [8] semantic distance is calculated and Curran Clark parser [9] used to parse each existing caption. The closest caption will be considered as the generated caption for the query image. Later deep learning technique is used for image captioning [12]. Retrieval based methods are effective for embedding and ranking but to generate the caption neural network based methodology is more effective [26]. Dependency tree recursive neural network is used to make sentences as composition vector [11]. To match the features into the global space max-margin technique is used. The performance of the network is being improved with time [26].

Retrieval and template base technique has few limitations such as generated sentence are dependent on the predefined template pool. The only advantage of these methods is grammatically correct caption. The resulting caption may not be important for the image scene. Deep neural network never believed in preexisting caption and there is no assumption about the generated sentence structure. These techniques are flexible in nature for the better caption generation. Deep learning-based techniques based on pure learning for sentence generation. Kiros suggested a neural network model for image captioning in which log-bilinear language is used. Mao [13] used the RNN based model to generate image caption. Pengfei Liu at el. [26] used CNN-CNN for image captioning.

III. Methodology

Our model mainly has three part.

1. Feature are extracted using the CNN.
2. Bahdanau Attention is used to form weighted matrices.
3. GRU is used to generate the sentence from those weighted features.

Block Diagram of the Proposed model is represented in Figure 1. Our model model consists of CNN (VGG16) which is used to encode the images. VGG16 model pre-trained on image-net data set [21]. GRU and attention mechanism is used as decoder. Network Architecture of our model represented in Figure (2).

A. Convolutional Neural Network

CNN can be used for object identification in a self-driving car, article recognition, picture titling etc. CNN consists of four basic layers. We can use these layers according to our requirements. These layers are Convolution, Rectified Linear Unit, Pooling and Fully Connected Layer. These layers are building blocks of the CNN. It is a feed-forward neural network. Each layer perform the convolution operation and element-wise summation as denoted in the equation (1).

$$F(x) = \sum_{n=1}^{N} G_k * (x) + b_k \qquad (1)$$

in the equation (1), F(x) denote the affine operation on the x, k is the total number of channel. b_k is the bias for k^{th} channel. To achieve non-linearity RELU function is used in eq. (2).

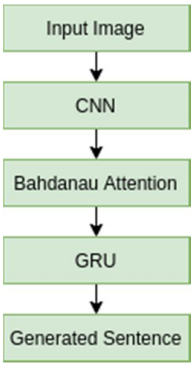

Figure 1: Block Diagram

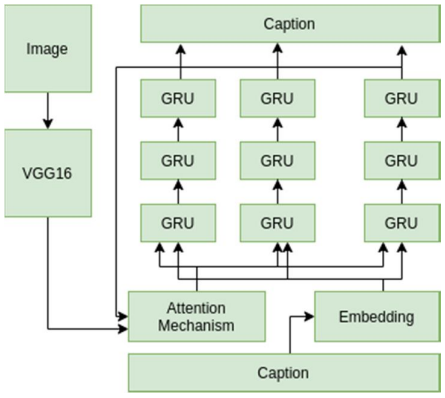

Figure 2: Network Architecture

$$R(x) = max(0, F(x)) \quad (2)$$

By combining the equations (1) and (2), generated equation (3).

$$F(x) = max(0, \sum_{k=1}^{N} G_k * x + b_k) \quad (3)$$

Where F(x) denotes the total units of CNN.

B. Gated Recurrent Unit

GRU enable the RNN network to capture the dependencies more accurately at different time scales [15]. GRU has gated units similar to LSTM which module information flow inside the cell. Current activation H_t^j at time t is the linear interpolation of candidate activation $\tilde{H}^t_j$ and previous activation H^j_{t-1} .

$$H_t^j = (1 - z_t^j)H_{t-1}^j + z_t^j H_t^j \quad (4)$$

where z_t^j is the update gate. It is like weight that decide how much to update its content or activation. Update gate can be calculate by using the equation (5).

$$z_t^j = \sigma(W_z x_t + U_z H_{t-1}) \quad (5)$$

Calculating linear sum between the existing state and new computed state is same in both LSTM and GRU unit. $\tilde{H}^t_j$ in equation (6) denote the candidate activation. It is same as the basic RNN cell.

$$\tilde{H}_t^j = tanh\left(W_x t + U(r_t \odot H_{t-1}^j)\right) \quad (6)$$

Where circled dot operator is element wise multiplication and r_t is a set of reset gates. Whenever r_t^j equal to zero than reset gate work as it is reading the initial set of pixel from input image. By utilizing this strategy it forgot the previous computed cell value. Reset gate r_t^j in equation (7) is calculated similar to update gate.

$$r_t^j = \sigma(W_r x_t + U_r H_{t-1}) \quad (7)$$

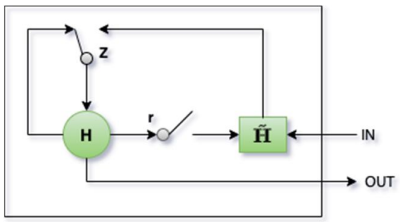

Fig. 3: Gated Recurrent Unit

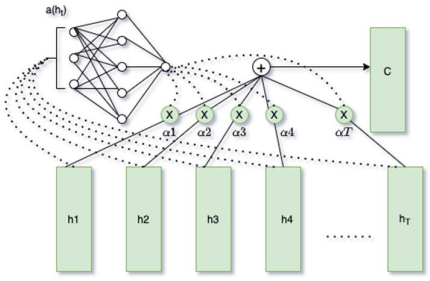

Fig. 4 Bahdanau Attention

C. Bahdanau Attention Mechanism

In general, the performance of vanilla RNN degrades when length of sequence is greater than 4-5. This is because RNN faces vanishing and exploding gradient problem. The performance of vanilla RNN to learn sequence ranging from 10-15 is increased with the support of LSTM and GRU [15].

To attenuate this issue Bahdanau et al. [28] suggested an approach in 2014, which uses attention mechanism. This mechanism learns from sum of past outputs by forming additional layer. This helps in determining which features the network wants to maintain. RNN can learn long sequences using the attention mechanism. To get the output for current cell attention mechanism uses the last state and the weighted mix of all states. Attention weights is used to retain the object for the next cell. It allow which input should be chosen for next cell from each previous input that is represented in figure (4). So it allows the network to catch the most important objects for the current cell. For a hidden unit h_t, context vector C_t is calculated as the weighted sum of state sequence.

$$C_t = \sum_{j=1}^{T} \alpha_{tj} h_j \quad (8)$$

where T denotes the number of time steps, $alpha_{tj}$ denotes weight at time step t for h_j. This context vector is then used to process a new state sequence $s_,$ where s_t depend upon past state sequence s_{t-1} and context vector C_t .

$$e_{tj} = F\left(\left(s_{t-1}, h_j\right)\right) \quad (9)$$

where, F is the learned function. It calculate the significance value of hidden unit h_j using the past state sequence s_{t-1}. The weight $alpha_{tj}$ can be calculated as given in equation (10).

$$\alpha_{tj} = \frac{exp\left(score\left(e_{tj}, h_s\right)\right)}{\sum_{k=1}^{T} exp(e_{tk}, h_s)} \quad (10)$$

IV. EXPERIMENT

A. Data set

We have trained our model on the Microsoft Common Objects in Context (MSCOCO) data set [22]. It contains about half million images and are divided in train, validation and test. Each image roughly contains 5 captions.

B. Evaluation Matrices

For the caption analysis many evaluation matrices exist. These are used to calculate the closeness of the sentences. Some well know evaluation matrices are BLEU, ROUGE, METEOR.

BLEU Score: The quality of generated text is assessed using BLEU score. BLEU score does not check the syntactic correctness of the sentences rather it matches each generated text with their respective reference texts that are composed by humans [17].

$$BLEU = min(1, G_L/R_L)\left(\prod_{i=1}^{4}(Precision_i)^{1/4}\right) \quad (11)$$

$$Precision = C_W/T_W \quad (12)$$

where G_L denotes the length of generated sentence, R_L denote the length of reference sentence/caption, C_W denote the correct words in generated sentences, T_W denote the total number of words in the generated sentence. Table (1) denote the BLEU scores on the test data-set.

TABLE I: BLEU SCORE ON TEST-DATA

Matrix Name	Calculated Value
BLEU-1	0.72
BLEU-2	0.55
BLEU-3	0.41
BLEU-4	0.31

ROGUE Score: We can also calculate the Rouge score to observe the quality of generated caption. Rouge match pair of words, sequences of words and n-gram with human composed reference summaries [30]. The calculated Rogue Score (CNN+GRU) is 65.

C. Performance of our model on Test Images

This section will show the performance of our model on the test images. Plotting the attention mechanism though which the caption is generated. It basically shows the words and the corresponding attention in image. The white space on the figure (6) represents higher weights (attention) which are given to those part of image using attention mechanism.

Actual Caption 1: A man riding a skateboard up the side of a ramp
Actual Caption 2: A man is doing a skateboarding trick on cement
Generated Caption: A man on a skateboard in the snow

Figure 5: Test Image

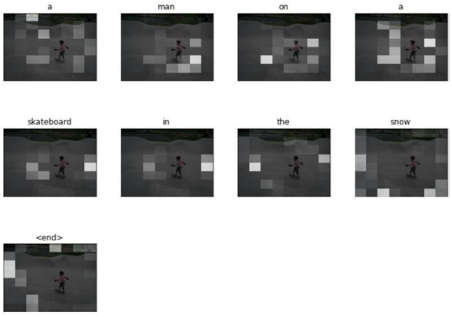

Figure 6: Attention Plot

Table II: Compare Result with Previous Models on MSCOCO dataset

Method	**BLEU-1**
M-RNN-AlexNet [23]	0.48
CNN-RNN [24]	0.69
Convolutional Decoder [25]	0.68
CNN-CNN [26]	0.71
Our CNN-GRU	0.72

We have compared our result with Multi-model Recurrent Neural Network [23], Convolutional Decoders For Image Captioning [25] and Convolutional Image Captioning [26] as shown in Table 2.

D. Loss Plot

We have trained our model for 50 epochs cross-entropy loss function is being used using ADAM optimiser [29]. It measures the quantitative loss at the each epoch. Loss value can be calculated as given in the equation (13).

$$Loss = \sum_{n=1}^{N} Y_{o=x,c} \, log\left(P_{x,c}\right) \qquad (13)$$

Where P is the probability of observation x in the class c. If y=1 it belong to correct class and if y=0 than it belong to incorrect class and N denotes the total number of classes.

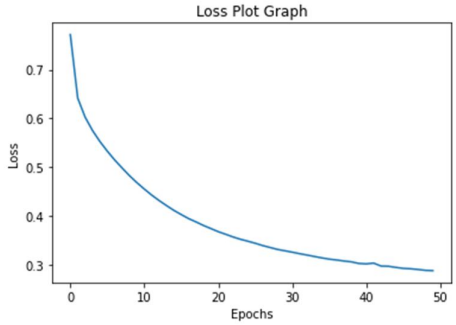

Figure 7: Loss Plo

V. Conclusion

We observed that our model used VGG16 to extract image features, Bhadanau Attention to adjust the weights and GRU is used to generate the the sentence. We have successfully trained and tested our model on MSCOCO dataset. The model is trained and tested on Google Colab. Our model has shown a signicant improvement on already exist CNN-CNN and CNN-RNN models. Calculated BLEU-1 score (0.72) is better as compared to previous neural network strategies. Further improvement can be done using bi-directional LSTM in encoding phase. We can also use hybrid loss function instead of cross-entropy loss.

References

[1] Ali Farhadi, Mohsen Hejrati, Mohammad Amin Sadeghi, Peter Young, Cyrus Rashtchian, Julia Hockenmaier, and David Forsyth. Every picture tells a story: Generating sentences from images. In Proceedings of the 11th European Conference on Computer Vision: Part IV, ECCV'10, pages 15–29, Berlin, Heidelberg, 2010. Springer-Verlag.

[2] Polina Kuznetsova, Vicente Ordonez, Alexander C. Berg, Tamara L. Berg, and Yejin Choi. Collective generation of natural image descriptions. In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers - Volume 1, ACL '12, pages 359–368, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics.

[3] Siming Li, Girish Kulkarni, Tamara L. Berg, Alexander C. Berg, and Yejin Choi. Composing simple image descriptions using web-scale n-grams. In Proceedings of the Fifteenth Conference on Computational Natural Language Learning, CoNLL '11, pages 220–228, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

[4] Xinlei Chen and C. Lawrence Zitnick. Learning a recurrent visual representation for image caption generation. CoRR, abs/1411.5654, 2014.

[5] Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, and Alan L. Yuille. Deep captioning with multimodal recurrent neural networks (m-RNN). CoRR, abs/1412.6632, 2014.

[6] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. CoRR, abs/1411.4555, 2014.

[7] Papineni, K. "BLEU: a method for automatic evaluation of MT." (2001).

[8] D. Lin, An information-theoretic definition of similarity, in: Proceedings of the Fifteenth International Conference on Machine Learning, pp. 296–304.

[9] J. Curran, S. Clark, J. Bos, Linguistically motivated large-scale nlp with cc and boxer, in: Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions, pp. 33–36.V. Ordonez, G. Kulkarni, T. L. Berg., Im2text: Describing images using 1 million captioned photographs, in: Advances in Neural Information Processing Systems, 2011, pp. 1143–1151.

[10] R. Socher, A. Karpathy, Q. V. Le, C. D. Manning, A. Y. Ng, Grounded compositional semantics for finding and describing images with sentences, TACL 2 (2014) 207–218.

[11] Kiros R, Salakhutdinov R, Zemel R (2014) Multimodal neural language models. In: International conference on machine learning, pp 595–603

[12] Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, Zhiheng Huang, and Alan Yuille. 2015. Deep captioning with multimodal recurrent neural networks (m-RNN). In International Conference on Learning Representations (ICLR).

[13] Cho, Kyunghyun; van Merrienboer, Bart; Gulcehre, Caglar; Bahdanau, Dzmitry; Bougares, Fethi; Schwenk, Holger; Bengio, Yoshua (2014). "Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation".

[14] Junyoung Chung Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling

[15] Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. In arXiv [cs.CV].

[16] K. Papineni, S. Roukos, T. Ward, and W. jing Zhu. Bleu: a method for automatic evaluation of machine translation. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), pages 311–318, 2002.

[17] Guadarrama, S., Krishnamoorthy, N., Malkarnenkar, G., Venugopalan, S., Mooney, R., Darrell, T., Saenko, K.: Youtube2text: Recognizing and Describing Arbitrary Activities Using Semantic Hierarchies and Zero-Shot Recognition. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2712–2719 (2013)

[18] R. Subash et al, Automatic Image Captioning Using Convolution Neural Networks and LSTM To cite this article: 2019

[19] Geetha V., Salvi S., Saini G., Yadav N., Singh Tomar R.P. (2021) "Follow Me: A Human Following Robot Using Wi-Fi Received Signal Strength Indicator" in Tuba M., Akashe S., Joshi A. (eds) ICT Systems and Sustainability. Advances in Intelligent Systems and Computing, vol 1270. Springer, SingaporeICT Systems and Sustainability. Advances in Intelligent Systems and Computing, vol 1270. Springer, Singapore.

[20] Deng, J. et al., 2009. Imagenet: A large-scale hierarchical image database. In 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255.

[21] Lin, T.-Y., Maire, M., Belongie, S. J., Bourdev, L. D., Girshick, R. B., Hays, J., Perona, P., Ramanan, D., Dollár, P. & Zitnick, C. L. (2014). Microsoft COCO: Common Objects in Context.. CoRR, abs/1405.0312.

[22] Mao, J., Xu, W., Yang, Y., Wang, J., Huang, Z., & Yuille, A. (2014). Deep captioning with multimodal recurrent neural networks (m-RNN). In arXiv [cs.CV].

[23] Liu, S., Bai, L., Hu, Y., & Wang, H. (2018). Image captioning based on deep neural networks. MATEC Web of Conferences, 232, 01052.
[24] A. Wang and A.B. Chan, “CNN+ CNN: Convolutional Decoders For Image Captioning”, arXiv preprint arXiv:1805.09019, 2018.
[25] J. Aneja, A. Deshpande, and A.G. Schwing, “Convolutional Image Captioning”, In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5561–5570, 2018
[26] Pengfei Liu ,Xipeng Qiu,Xuanjing Huan"Recurrent Neural Network for Text Classification with Multi-Task Learning" Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI-16)
[27] Jan K. Chorowski, Dzmitry Bahdanau, Dmitriy Serdyuk, Kyunghyun Cho, Yoshua Bengio "Attention-Based Models for Speech Recognition" in Advances in Neural Information Processing Systems 28 (NIPS 2015).
[28] Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. In arXiv [cs.LG].
[29] Tingting He, Jinguang Chen, Liang Ma, Zhuoming Gui, Fang Li, Wei Shao, Qian Wang "ROUGE-C: A fully automated evaluation method for multi-document summarization" in 2008 IEEE International Conference on Granular Computing. Eason, B. Noble, and I. N. Sneddon, “On certain integrals of Lipschitz-Hankel type involving products of Bessel functions,” *Phil. Trans. Roy. Soc. London*, vol. A247, pp. 529–551, April 1955.