SAKARYA UNIVERSITY JOURNAL OF COMPUTER AND INFORMATION SCIENCES
VOL. 4, NO. 2, AUGUST 2021

Research Article

Deep Gated Recurrent Unit for Smartphone-Based Image Captioning

Volkan Kılıç[1]
[1]Department of Electrical and Electronics Engineering, Izmir Katip Celebi University, Izmir, 35620, Turkey;
volkan.kilic@ikcu.edu.tr; +90 232 329 35 35/3881

Received 22 January 2021; Revised 11 May 2021; Accepted 13 May 2021; Published online 31 August 2021

Abstract

Expressing the visual content of an image in natural language form has gained relevance due to technological and algorithmic advances together with improved computational processing capacity. Many smartphone applications for image captioning have been developed recently as built-in cameras provide advantages of easy-operation and portability, resulting in capturing an image whenever or wherever needed. Here, an encoder-decoder framework based new image captioning approach with a multi-layer gated recurrent unit is proposed. The Inception-v3 convolutional neural network is employed in the encoder due to its capability of more feature extraction from small regions. The proposed recurrent neural network-based decoder utilizes these features in the multi-layer gated recurrent unit to produce a natural language expression word-by-word. Experimental evaluations on the MSCOCO dataset demonstrate that our proposed approach has the advantage over existing approaches consistently across different evaluation metrics. With the integration of the proposed approach to our custom-designed Android application, named "*VirtualEye+*", it has great potential to implement image captioning in daily routine.

Keywords: artificial intelligence, natural language processing, image captioning, Android

1. Introduction

The problem of image captioning has received much attention from the computer vision (CV) and natural language processing (NLP) communities in recent decades due to its potential applications including image indexing or retrieval, virtual assistants for visually impaired people [1, 2]. Image captioning needs a higher level of image understanding beyond object detection and classification to identify the objects and actions which plays a critical role to generate expressions for an image in the form of a natural language with proper linguistic properties. Early efforts in image captioning often use either template-based [3-7] methods or retrieval-based [8-11] methods. The template-based methods employ image information such as objects, scenes and attributes to generate a meaningful caption using the most relevant words from sentence templates. The captions have constant length and highly sensitive to the performance of the object detector which leads to simple sentences with a tendency to deviate from the ground truth captions. To address these problems, the visual information of the input image was employed in the retrieval-based methods to match ground truth captions of the most likely images from the retrieval library. The matched ground truth captions are collected to get more flexible and semantically richer captions. The main drawback in the retrieval-based methods is that the generated captions may be misleading if similar images are not contained in the retrieval library.

These drawbacks in the template-based and retrieval-based methods have been overcome with neural network based methodologies which combine convolutional neural network (CNN) and recurrent neural network (RNN) [12, 13]. The problem of image captioning is formulated as a translation problem [14] inspired by machine translation [15, 16]. The CNN and RNN methods are used under the encoder-decoder framework [17-19] which leads to employing deep networks in image captioning. The encoder consists of deep CNNs used to extract visual information from an image. Recently, advanced CNN architectures including NASNetLarge [20], Xception [21] and Inception-v3 [22] have emerged which show promising performance under various encoder designs. The RNN-based decoder converts the extracted information by CNN-based encoders into natural language captions word-by-word. Conventional RNNs, however, have vanishing and exploding gradient problems which prevents to employ of sufficiently long-term temporal dependencies [13, 23, 24]. To address these issues, long short-term memory (LSTM) [25] and gated recurrent unit (GRU) [26] networks are proposed. The

To Cite This Article: V. Kilic, "Deep Gated Recurrent Unit for Smartphone-based Image Captioning," Sakarya University Journal of Computer and Information Sciences, vol. 4, no. 2, pp. 181–191, 2021. doi: 10.35377/saucis.04.02.866409

LSTM uses a memory cell to store information for long periods of time in memory while GRU keeps the flow of information without additional memory cells. Chen et al. [27] proposed a style-factual image caption generator that uses the encoder-decoder framework with LSTM and they injected the style-factual features into the decoder. The factual representation is obtained by utilizing an adaptive learning approach. You et al. [28] presented an image caption generator where they employed the encoder-decoder framework with LSTM. Term generator and language generator were proposed in [29] where the CNN features of images are fed with terms from reference captions into the initial hidden state of the GRU. The term generator produces words which are the input of the language generator for the final caption.

RNN-based decoders generally process visual information from the encoder under two approaches [30]. The first approach includes direct feeding from the encoder into the RNN while the latter approach uses one additional layer before the RNN. These approaches can be sorted as init-inject, pre-inject, par-inject and merge architectures [30]. The init-inject receives the visual information as an image vector to feed to the initial hidden state of the RNN [31]. The representative study in this category is the scene graph [31] which detects the objects and extracts attributes to feed into the initial hidden state vector of the LSTM based RNN together with the CNN image features. In pre-inject architecture, the RNN takes the image vector as a first input [32-35] while the par-inject architecture employs the image vector and word vectors of the caption prefix in parallel as an input to the RNN [14, 36]. The merge-architecture employs the image after the RNN generates the caption prefix instead of feeding the image vector directly to the RNN [12, 37, 38]. The comparison of these architectures was reported in [30] that the init-inject shows superior performance in terms of generation and retrieval measures.

In this study, a neural encoder-decoder framework based new image captioning model is proposed which encodes the images with Inception-v3 CNN to generate captions using multi-layer GRU based RNN decoder under init-inject architecture. Among the comparison of NASNetLarge, Xception and Inception-v3 on the proposed image captioning model, Inception-v3 outperforms the others in terms of encoding visual information. In RNN based decoder, GRU is employed because of its computational efficiency and simplicity as it includes one hidden state vector while LSTM is operated with two state vectors consists of hidden and cell states [39]. Moreover, LSTM has input, forget and output gates while GRU uses only two gates: update and reset. In terms of compatibility with init-inject architecture, GRU with one hidden state vector also offers the best choice [30]. The motivation behind using multiple layers in GRU is to utilize the most relevant information in the unit which improves the ability of the decoder in utilizing visual information, so that an enhanced prediction model for caption generation is provided [40]. The proposed model is then compared with other existing approaches for caption generation, using performance metrics such as BLEUn (with n = 1, …, 4) [41], METEOR [42], CIDEr [43], and ROUGEL [44]. Typically, BLEUn is a type of n-grams precision measure as BLEU3 means 3-grams. METEOR is the harmonic mean of the precision and recall score for unigrams while CIDEr measures the similarity of n-grams using weighted cosine function, and ROUGEL is a score of longest common subsequence [45].

The rest of the paper is organized as follows: Section 2 introduces the theoretical foundations for the encoder and decoder framework. The proposed image captioning approach is given in Section 3. Experimental results performed on the MSCOCO dataset and performance comparison of the approaches are discussed in Section 4, followed by the conclusions.

2. Encoder-Decoder Framework for Image Captioning

The theoretical foundation of our proposed approach for image captioning based on the encoder-decoder framework is described in this section. The CNN architectures employed to obtain visual features and attributes of the image are introduced before the RNN-based decoders which are used to produce the image captions.

2.1 Encoder

Encoding an image means converting image data into a feature vector that contains the image information. Conventional encoder designs are based on a CNN due to its ability to deal with high dimensional data and remarkable feature extraction capability. In CNN, there are convolutional, pooling, and fully connected layers. In the convolutional layer, a filter is convolved with an image to create a feature or activation map which contains the detected features in the image. The pooling layer is a sampling layer that gradually decreases the spatial size of the feature map, resulting in a reduced number of features and computational complexity [46]. The fully connected layer produces the final decision based on all input from the previous layers.

Image captioning requires advanced computer vision techniques for image analysis and feature extraction. Deep CNN architectures best fit the requirement of convenient feature extraction that improves the quality of the captions. Therefore, pre-trained deep CNN architectures are used in this study such as NASNetLarge, Xception, and Inception-v3 in the encoder part.

Inception-v3 is a deep CNN, which is consists of 42-layers of convolutional, pooling, and fully connected layers. This architecture procured second place in ILSVRC 2015. Xception is a novel deep CNN obtained by modifying Inception-v3 with depth-wise sectional convolutions. Therefore, Xception architecture surpassed Inception-v3 on the ImageNet dataset. NASNetLarge (Neural Architecture Search Network) is a constructed CNN architecture designed using reinforcement learning on the CIFAR-10 dataset [47]. The ImageNet dataset is used in training of the architecture which leads to state-of-the-art performance.

The encoder extracts a high-level feature from the image using the convolutional and pooling layers of the CNN architectures. Then, the features are fed into the decoder for caption generation.

2.2 Decoder

A decoder produces words to describe an image with semantically meaningful sentences by using feature representation. Decoders are mostly designed based on RNN as it is capable of storing parts of the inputs and use them to generate meaningful captions.

RNN is a type of deep network that uses its internal state to process input sequences, which makes it suitable for sequential applications including speech recognition and image captioning [46, 48]. RNN calculates each output employing the same function over each instance of the sequence repeatedly. RNN consists of a hidden state and an optional output which operate on the input sequence. The current hidden state has been computed by taking the current input with the hidden state for the former time step using a nonlinear activation function which leads to an update of the output at each time step. The motivation behind using the RNN relies on the generalization of the solution with respect to time and its capability to deal with sequences which a classical deep learning architecture can not be applied directly. However, RNN suffers from the problem of vanishing and exploding gradients. Therefore, it cannot maintain long term dependencies. This problem is addressed by employing GRU which is a type of RNNs with a gating mechanism.

Conventional GRU with a hidden state, update and reset gates is depicted in Figure 1.

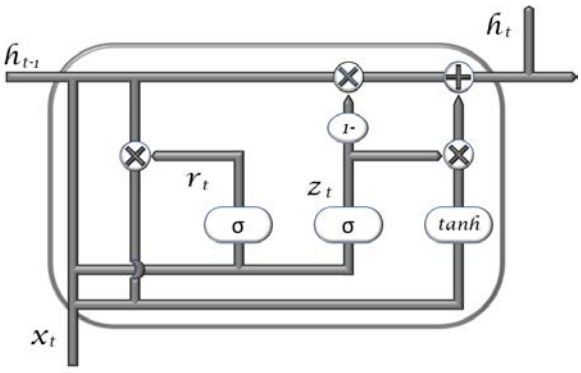

Figure 1 Gated Recurrent Unit

In GRU, the flow of information is maintained with following equations [49]:

$$r_t = \sigma(W_{xr}x_t + U_{hr}h_{t-1}) \quad (1)$$

$$z_t = \sigma(W_{xz}x_t + U_{hz}h_{t-1}) \quad (2)$$

$$u_t = tanh(W_{xu}x_t + U_{hu}(r_t \odot h_{t-1})) \quad (3)$$

$$h_t = (1 - z_t)h_{t-1} + z_t u_t \quad (4)$$

where x_t and h_t are the input and output vectors, respectively. Reset gate vector is defined with r_t and z_t denotes the update gate. The tangent hyperbolic activation function is denoted with tanh while σ is the sigmoid functions. Parameters are defined with W and U matrices, and ⊙ defines element-wise multiplication. The vanishing gradient problem is handled by a gating mechanism in the GRU while the exploding gradient is addressed with gradient clipping strategy [50]. The input is taken from the previous layer at each time step which provides to configure the GRU with multiple layers, resulting in outstanding performance compared to the conventional RNN-based architectures on many NLP tasks, including language modeling [51, 52].

3. Proposed Image Captioning Approach

This section presents a new approach to improve the image captions by introducing multi-layer GRU into an RNN-based decoder. After the proposed approach is described in the next subsection, our custom-designed Android application which runs the proposed approach under a user-friendly interface is demonstrated.

3.1 Deep Gated Recurrent Unit for Image Captioning

The encoder-decoder framework comprises CV and NLP algorithms in the encoder and decoder, respectively. CNN based encoders implement the CV tasks by extracting feature representation of an image while RNN-based decoders translate this representation into natural language captions. Image features and linguistic features can be deployed in RNN using different types of architectures, such as init-inject, pre-inject, par-inject and merge architectures [30]. The feature vector of the image has been employed as RNN initial hidden state vector in init-inject architecture. The image feature vector and the hidden state vector of the RNN should be the same size to meet the requirement of an early binding architecture which helps the RNN to change the image representation. It is noted in [30] that the init-inject outperforms the others in terms of generation and retrieval measures. In this study, a new deep GRU design is proposed in an RNN-based decoder under init-inject architecture for natural language descriptions of the image.

The queried image features are processed to generate a caption by the decoder involving an embedding layer, GRUs, and a fully connected (FC) layer. The proposed RNN-based decoder with multi-layer GRU is given in Figure 2. The GRU learns how to process image features and vectors to generate the most meaningful attributions. The embedding layer represents words as meaningful vectors. The fully connected layer predicts the most applicable word corresponding to the attributions.

As CNNs are not capable of handling word sequences, the conversion of words to vectors is needed to process in the RNN. Word embedding is the common approach to obtain vectors that contain semantics of the corresponding words. Here, the words were indexed into integer-tokens and converted into 128-sized float arrays by using an embedding layer. The embedding layer was trained along with the network to capture the more compact features of the words.

The captions were considered as time-series data of words and GRUs were utilized to learn the connection between the words in a caption. The four-layer GRU was constructed by combining the four individual RNN with their initial states. All features from the encoder were split into four equal-sized vectors, and each vector was fed to the initial state of GRU layers sequentially. Time series data from the output of the embedding layer was employed as input for the GRU layers. The correct token is predicted by a fully connected layer at the end of the decoder for semantically more meaningful captions.

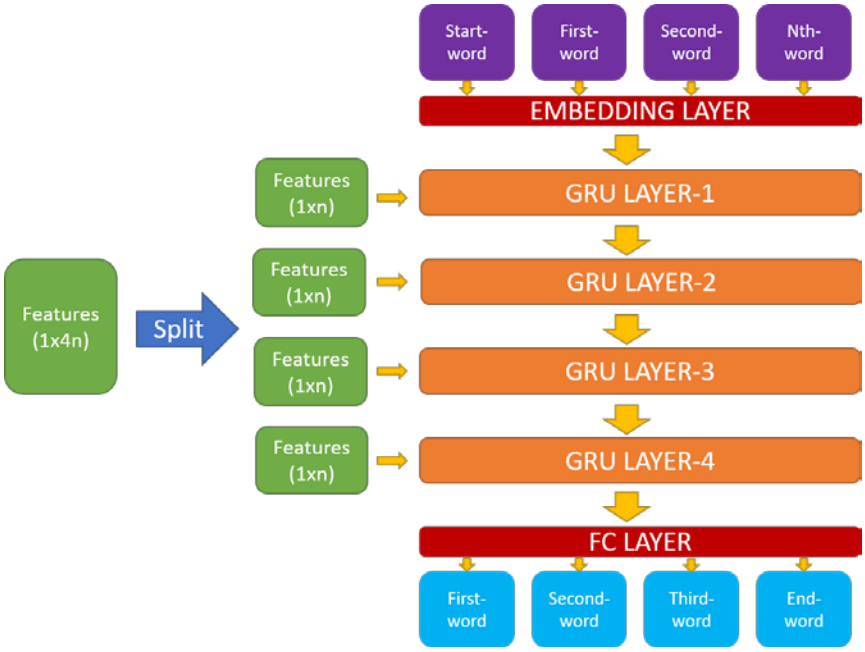

Figure 2 Multi-layer GRU based Decoder

3.2 Smartphone Application: *Virtual Eye+*

Our previous custom-design Android application *Virtual Eye* [53] was upgraded with new features and named it *Virtual Eye+* which provides more simplicity and easy operation.

The improvements on the application can be sorted into three categories; user interface, server, and cloud communication. First, the user interface was improved to increase the capability of easy-operation and portability, so that it can be benefited whenever captioning is needed. The homepage provides a manual that introduces the application for image captioning. In addition, “choose file” and “generate caption” buttons are given on the home screen. An image can be chosen from the gallery by tapping the “choose file” button. When the “generate caption” button is tapped, the application sends the selected image to the server and waits for the caption. The received caption is displayed on the homepage under the input image and it can be audible when the input image is tapped. Moreover, an in-app camera is provided where the user can access it by scrolling left from the homepage. An image can be captured by tapping anywhere on the screen and then, *Virtual Eye+* automatically sends the photo to the server and scrolls back to the homepage. The generated caption on the server is English as a default. However, it is possible to change language from the settings of the *Virtual Eye+* and so that the English caption can be translated to the smartphone display language.

Second, the server was set up on the Kivy user interface which generates a caption for the input image. If the user has requested a “non-English” caption, the caption is translated using the translate API from google before sending it to the user. Lastly, cloud communication is fulfilled under the Firebase cloud service which allows fast and robust communication between Python-based server and Java-based Android application. When an image is uploaded to the Firebase storage, its download link is activated which invokes the server to download the image via the link. The overall system is demonstrated in Figure 3.

4. Experimental Evaluations

4.1 Dataset and Performance Metrics

In order to evaluate the proposed captioning approach and compare it with existing approaches, a dataset including a large number of images with reference captions is required.

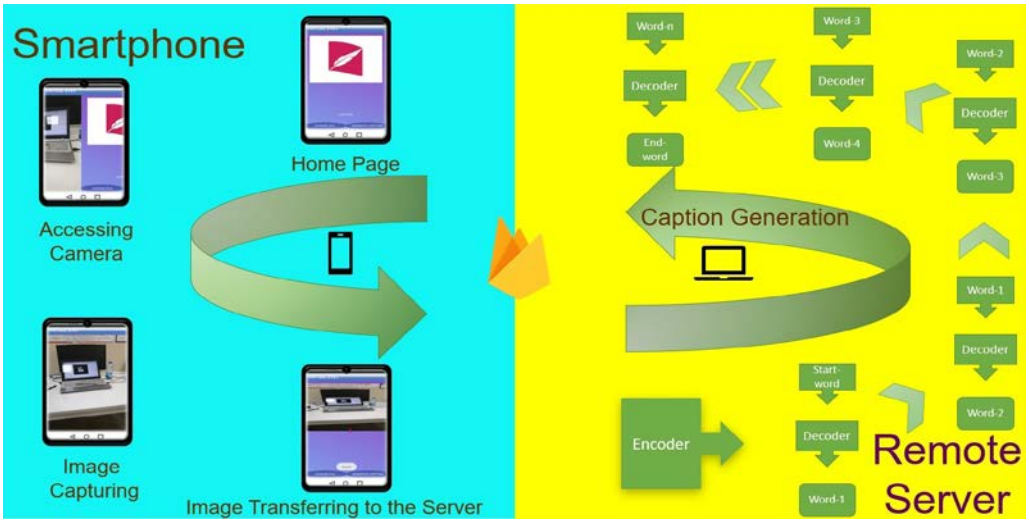

Figure 3 The Working Principle of Virtual Eye+

Apart from MSCOCO [54], the suitability of several other publicly available captioning datasets such as Flickr [55], and VizWiz-Captions [56] have been investigated and concluded that only MSCOCO is suitable for the evaluation of our proposed approach. Flickr offers two types of sub-datasets named Flickr8k and Flickr30k. Flickr8k includes 8000 images consisting of 6000 training, 1000 test and 1000 validation images. In Flickr30k, there are 29783 training, 1000 test, 1000 validation and a total of 31783 images. On the other hand, the VizWiz dataset includes 23431 training, 8000 test and 7750 validation images captured by people who are blind. The MSCOCO dataset has more than 120,000 images along with 5 captions for each image. This dataset dominates the image captioning studies with its grammatically and semantically correct captions with many diverse images which lead us to choose the MSCOCO dataset.

To analyze the performance of the compared approaches, several metrics including BLEU-n, ROUGE-L, and CIDEr are employed. BLEU-n compares a machine-generated caption by n-gram pairs to the human-generated ground truth captions [41]. Searching for pairs on short captions may arise a problem as getting a higher score, even the result is not correct. Hence, BLEU-n applies the brevity penalty to overcome the problem. ROUGE-L determines the longest co-occurring in word-sequence n-grams by itself [44]. CIDEr metric calculates the resemblance of the generated caption to a series of ground truth captions [43]. These metric results highly correlate with human judgments. The higher results are obtained on these metrics, the better captions are produced by the image caption generator. Among these three metrics, CIDEr is the only one designed for the image captioning problem while others are originally derived for machine translation. Therefore, methods are sorted based on CIDEr metric in the next.

In the experiment with the decoder part, the RMSprop optimizer was employed to update the parameters. The learning rate was set to be $1x10^{-3}$. The loss function was chosen as cross-entropy loss with the combination of the negative log likelihood loss and the logarithmic softmax function. The size of the feature vector was 2048. The vocabulary size, embedding size and hidden size of GRU layers were set to be 10.000, 128 and 512, respectively. The input size of the FC layer is equal to the hidden size of GRU whereas the output size is the vocabulary size. An activation function *linear* was employed in the FC layer.

4.2 Results and Discussion

The proposed deep GRU based decoder firstly was tested with three different encoders in order to find the best CNN architecture compatible with our multi-layer GRU design. The Inception-v3, Xception, and NASNetLarge were evaluated under BLEU-n, ROUGE-L, and CIDEr metrics, and the scores were given in Table 1 which demonstrates that the proposed captioning approach is consistently better with the Inception-v3 CNN encoder.

Table 1 Performance of Deep GRU-Based Decoder with Three Different CNN Encoders

Encoders	**BLEU-1**	**BLEU-2**	**BLEU-3**	**BLEU-4**	**ROGUE-L**	**CIDEr**
NASNetLarge	0.644	0.455	0.312	0.213	0.473	0.748
Xception	0.650	0.466	0.324	0.225	0.479	0.770
Inception-v3	**0.652**	**0.470**	**0.330**	**0.232**	**0.484**	**0.775**

(a)

(b)

Figure 4 Sample Images from MSCOCO Dataset

To demonstrate the generated captions, two images were selected from MSCOCO dataset given in Figure 4 while the ground truth and generated captions were given in Table 2. If the generated captions with Inception-v3, Xception and NASNetLarge are compared, it can be seen that the caption from Inception-v3 is semantically more meaningful and closer to the ground truth captions.

Table 2 Ground Truth and Generated Captions for the Images of Figure 4

Ground Truth Captions for Figure 4 (a):	**Ground Truth Captions for Figure 4 (b):**
A big burly grizzly bear is show with grass in the background.	A large white bowl of many green apples.
The large brown bear has a black nose.	A white bowl of green granny smith apples.
Closeup of a brown bear sitting in a grassy area.	A white bowl filled with green Granny Smith apples.
A large bear that is sitting on grass.	A bowl filled with many shiny green apples.
A close up picture of a brown bear's face.	A bowl full of fresh green apples are kept.
Generated captions:	**Generated captions:**
Inception-v3: **a grizzly bear is laying down on a green field.**	**Inception-v3**: **a pile of green apples on a table.**
Xception: a large brown bear laying on top of a tree.	Xception: a bowl of fruit sitting on a table.
NASNetLarge: a grizzly bear sitting on a rock with a grassy field behind him.	NASNetLarge: a plate topped with a green apple and a green apple.

The Inception-v3 based image caption generator has been integrated into the *VirtualEye+* as its outstanding score compared to the other CNN encoders. The caption generation time is about 10 seconds depending on the internet connection for the smartphone application. The proposed approach is also compared with those of Chen et al. [27], You et al. [28], Xu et al. [31] and Mathews et al. [31] under available metrics reported on their papers. In each column, the highest score is indicated with bold fonts and the approaches are sorted based on the CIDEr metric. Even though the results of [31] are slightly better than ours in terms of the BLEUn metrics, the proposed approach outperforms the others with respect to ROGUE-L and CIDEr metrics.

Table 3 Performance Metric Results

Methods	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROGUE-L	CIDEr
[27]	0.505	0.308	0.191	0.121	-	0.600
[28]	0.510	0.322	0.207	0.136	0.390	0.654
[31]	**0.664**	**0.482**	**0.337**	0.233	**0.484**	0.689
[29]	0.653	-	-	**0.238**	-	0.769
Proposed Approach	0.652	0.470	0.330	0.232	**0.484**	**0.775**

5. Conclusion

Herein, we proposed a new image captioning approach based on the Inception-v3 CNN encoder and deep GRU decoder. The deep decoder design has been investigated in natural language expressions of images with multi-layer sequential GRUs. The proposed approach was trained on the MSCOCO dataset. The results showed that captioning performance was significantly improved with a multi-layer GRU based decoder and outperformed the state-of-the-art approaches. Then, the proposed approach was integrated with the *VirtualEye+* Android application to offer easy-operation of image captioning under a user-friendly interface. An image can be taken either from the gallery or camera to transfer to the remote server via the Firebase. The remote server runs our proposed approach to generate captions which can be transferred back to the application either in English or any other language that the user requests. In addition, *VirtualEye+* provides narrator options to allow the user to hear the generated captions. In the proposed approach, CNN encoder and RNN decoder could be embedded in the smartphone application to reduce the caption generating time which could be interesting directions for future work.

References

[1] B. Makav and V. Kılıç, "A New Image Captioning Approach for Visually Impaired People," in *11th International Conference on Electrical and Electronics Engineering*, 2019, pp. 945-949: IEEE.

[2] B. Makav and V. Kılıç, "Smartphone-based Image Captioning for Visually and Hearing Impaired," in *11th International Conference on Electrical and Electronics Engineering*, 2019, pp. 950-953: IEEE.

[3] G. Kulkarni *et al.*, "Baby talk: Understanding and generating image descriptions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2011, pp. 1601-1608.

[4] M. Mitchell *et al.*, "Midge: Generating image descriptions from computer vision detections," in *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, 2012, pp. 747-756: Association for Computational Linguistics.

[5] D. Elliott and F. Keller, "Image description using visual dependency representations," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2013, pp. 1292-1302.

[6] X. Zhang, X. Wang, X. Tang, H. Zhou, and C. Li, "Description generation for remote sensing images using attribute attention mechanism," *Remote Sensing*, vol. 11, no. 6, p. 612, 2019.

[7] H. Fang *et al.*, "From captions to visual concepts and back," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1473-1482.

[8] R. Mason and E. Charniak, "Nonparametric method for data-driven image captioning," in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 2014, pp. 592-598.

[9] P. Kuznetsova, V. Ordonez, T. L. Berg, and Y. Choi, "Treetalk: Composition and compression of trees for image descriptions," *Transactions of the Association for Computational Linguistics*, vol. 2, pp. 351-362, 2014.

[10] R. Socher, A. Karpathy, Q. V. Le, C. D. Manning, and A. Y. Ng, "Grounded compositional semantics for finding and describing images with sentences," *Transactions of the Association for Computational Linguistics*, vol. 2, pp. 207-218, 2014.

[11] M. Yang *et al.*, "An Ensemble of Generation-and Retrieval-Based Image Captioning With Dual Generator Generative Adversarial Network," *IEEE Transactions on Image Processing*, vol. 29, pp. 9627-9640, 2020.

[12] J. Mao, W. Xu, Y. Yang, J. Wang, Z. Huang, and A. Yuille, "Deep captioning with multimodal recurrent neural networks (m-rnn)," *arXiv preprint arXiv:1412.6632*, pp. 1-17, 2015.

[13] A. Oluwasammi *et al.*, "Features to Text: A Comprehensive Survey of Deep Learning on Semantic Segmentation and Image Captioning," *Complexity*, vol. 2021, 2021.

[14] J. Donahue *et al.*, "Long-term recurrent convolutional networks for visual recognition and description," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 2625-2634.

[15] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:*. 2014.

[16] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Advances in Neural Information Processing Systems*, 2014, pp. 3104-3112.

[17] R. Kiros, R. Salakhutdinov, and R. S. Zemel, "Unifying visual-semantic embeddings with multimodal neural language models," *arXiv preprint arXiv:*. 2014.

[18] D. W. Otter, J. R. Medina, and J. K. Kalita, "A survey of the usages of deep learning for natural language processing," *IEEE Transactions on Neural Networks Learning Systems*, 2020.

[19] S. Kalra and A. Leekha, "Survey of convolutional neural networks for image captioning," *Journal of Information Optimization Sciences*, vol. 41, no. 1, pp. 239-260, 2020.

[20] B. Zoph, V. Vasudevan, J. Shlens, and Q. V. Le, "Learning transferable architectures for scalable image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8697-8710.

[21] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1251-1258.

[22] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2818-2826.

[23] M. Z. Hossain, F. Sohel, M. F. Shiratuddin, and H. Laga, "A comprehensive survey of deep learning for image captioning," *ACM Computing Surveys*, vol. 51, no. 6, pp. 1-36, 2019.

[24] H. Wang, Y. Zhang, and X. Yu, "An Overview of Image Caption Generation Methods," *Computational Intelligence Neuroscience*, vol. 2020, 2020.

[25] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735-1780, 1997.

[26] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *arXiv preprint arXiv:*. 2014.

[27] T. Chen *et al.*, "``Factual"or``Emotional": Stylized Image Captioning with Adaptive Learning and Attention," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 519-535.

[28] Q. You, H. Jin, and J. Luo, "Image captioning at will: A versatile scheme for effectively injecting sentiments into image descriptions," *arXiv preprint arXiv:.10121*, 2018.

[29] A. Mathews, L. Xie, and X. He, "SemStyle: Learning to Generate Stylised Image Captions Using Unaligned Text," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8591-8600.

[30] M. Tanti, A. Gatt, and K. P. Camilleri, "Where to put the image in an image caption generator," *Natural Language Engineering*, vol. 24, no. 3, pp. 467-489, 2018.

[31] N. Xu, A.-A. Liu, J. Liu, W. Nie, and Y. Su, "Scene graph captioner: Image captioning based on structural visual representation," *Journal of Visual Communication Image Representation*, vol. 58, pp. 477-485, 2019.

[32] O. Nina and A. Rodriguez, "Simplified LSTM unit and search space probability exploration for image description," in *2015 10th International Conference on Information, Communications and Signal Processing (ICICS)*, 2015, pp. 1-5: IEEE.

[33] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3156-3164.

[34] S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel, "Self-critical sequence training for image captioning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 7008-7024.

[35] J. Krause, J. Johnson, R. Krishna, and L. Fei-Fei, "A hierarchical approach for generating descriptive image paragraphs," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 317-325.

[36] T. Yao, Y. Pan, Y. Li, Z. Qiu, and T. Mei, "Boosting image captioning with attributes," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 4894-4902.

[37] J. Mao, X. Wei, Y. Yang, J. Wang, Z. Huang, and A. L. Yuille, "Learning like a child: Fast novel visual concept learning from sentence descriptions of images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 2533-2541.

[38] T. A. Praveen and J. A. A. Jothi, "Enhancing Image Caption Quality with Pre-post Image Injections," in *Advances in Machine Learning and Computational Intelligence*: Springer, 2021, pp. 805-812.

[39] H. Wang, H. Wang, and K. Xu, "Evolutionary Recurrent Neural Network for Image Captioning," *Neurocomputing*, 2020.

[40] Y. Tao, X. Wang, R.-V. Sánchez, S. Yang, and Y. Bai, "Spur gear fault diagnosis using a multilayer gated recurrent unit approach with vibration signal," *IEEE Access*, vol. 7, pp. 56880-56889, 2019.

[41] 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 of the Association for Computational Linguistics*, 2002, pp. 311-318.

[42] S. Banerjee and A. Lavie, "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments," in *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, 2005, pp. 65-72.

[43] R. Vedantam, C. Lawrence Zitnick, and D. Parikh, "Cider: Consensus-based image description evaluation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 4566-4575.

[44] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Text Summarization Branches Out*, 2004, pp. 74-81.

[45] K. Drossos, S. Lipping, and T. Virtanen, "Clotho: An Audio Captioning Dataset," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 736-740: IEEE.

[46] L. Zhang, S. Wang, and B. Liu, "Deep learning for sentiment analysis: A survey," *Journal of Wiley Interdisciplinary Reviews: Data Mining Knowledge Discovery*, vol. 8, no. 4, p. e1253, 2018.

[47] A. Krizhevsky and G. Hinton, "Learning multiple layers of features from tiny images," 2009.

[48] X. Li and X. Wu, "Constructing long short-term memory based deep recurrent neural networks for large vocabulary speech recognition," in *International Conference on Acoustics, Speech and Signal Processing*, 2015, pp. 4520-4524: IEEE.

[49] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *arXiv preprint arXiv:1412.3555*, 2014.

[50] S. Hochreiter, "The vanishing gradient problem during learning recurrent neural nets and problem solutions," *International Journal of Uncertainty, Fuzziness Knowledge-Based Systems*, vol. 6, no. 02, pp. 107-116, 1998.

[51] M. Sundermeyer, R. Schlüter, and H. Ney, "LSTM neural networks for language modeling," in *Thirteenth Annual Conference of the International Speech Communication Association*, 2012.

[52] R. Pascanu, T. Mikolov, and Y. Bengio, "On the difficulty of training recurrent neural networks," in *International Conference on Machine Learning*, 2013, pp. 1310-1318: PMLR.

[53] Ö. Çaylı, B. Makav, V. Kılıç, and A. Onan, "Mobile Application Based Automatic Caption Generation for Visually Impaired," in *International Conference on Intelligent and Fuzzy Systems*, 2020, pp. 1532-1539: Springer.

[54] T.-Y. Lin *et al.*, "Microsoft coco: Common objects in context," in *European Conference on Computer Vision*, 2014, pp. 740-755: Springer.

[55] 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," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 2641-2649.

[56] D. Gurari, Y. Zhao, M. Zhang, and N. Bhattacharya, "Captioning Images Taken by People Who Are Blind," *arXiv preprint arXiv:.08565*, 2020.