Avrupa Bilim ve Teknoloji Dergisi
Sayı 35, S. 610-615, Nisan 2022

Araştırma Makalesi

www.ejosat.com ISSN:2148-2683

European Journal of Science and Technology
No. 35, pp. 610-615, April 2022

Research Article

Resnet based Deep Gated Recurrent Unit for Image Captioning on Smartphone

Betül Uslu[1*], Özkan Çaylı[1], Volkan Kılıç[1], Aytuğ Onan[2]

[1] Izmir Katip Celebi University, Faculty of Engineering, Departmant of Electrical and Electronics, İzmir, Turkey, (ORCID: 0000-0003-1868-9670, 0000-0002-3389-3867, 0000-0002-3164-1981), betuluslu5u5@gmail.com, ozkan.cayli@ikcu.edu.tr, volkan.kilic@ikcu.edu.tr
[2] Izmir Katip Celebi University, Faculty of Engineering, Departmant of Computer, İzmir, Turkey, (ORCID: 0000-0002-9434-5880), aytug.onan@ikcu.edu.tr

(First received 21 February 2022 and in final form 30 April 2022)

(**DOI:** 10.31590/ejosat.1107035)

ATIF/REFERENCE: Uslu, B., Çaylı, Ö., Kılıç, V., & Onan, A. (2022). Resnet based Deep Gated Recurrent Unit for Image Captioning on Smartphone. *European Journal of Science and Technology*, (35), 610-615.

Abstract

Image captioning aims at generating grammatically and semantically acceptable natural language sentences for visual contents. Gated recurrent units (GRU) based approaches have recently attracted much attention due to their performance in caption generation. Challenges with GRU are to deal with vanishing gradient problems and modulation of the most relevant information flow in deep networks. In this paper, we propose a resnet-based deep GRU approach to overcome the vanishing gradient problem with residual connections while the most relevant information is ensured to flow using multiple layers of GRU. Residual connections are employed between consecutive layers of deep GRU, which improves the gradient flow from lower to upper layers. Experimental investigations on the publicly available MSCOCO dataset prove that the proposed approach achieves comparable performance with some state-of-the-art approaches. Moreover, the approach is embedded into our custom-designed Android application, *CaptionEye*, which offers the ability to generate captions without an internet connection under a voice user interface.

Keywords: Gated Recurrent Unit, Residual Connection, Image Captioning, Android Application.

Resnet Tabanlı Derin Geçitli Tekrarlayan Birim ile Akıllı Telefonda Görüntü Altyazılama

Öz

Görüntü altyazılama, görsel içerikler için dilbilgisel ve anlamsal olarak uygun doğal dil cümleleri oluşturmayı amaçlamaktadır. Geçitli tekrarlayan birim (GRU) tabanlı yaklaşımlar, son zamanlarda altyazı oluşturmadaki performanslarından dolayı büyük ilgi görmektedir. Kaybolan gradyan problemi ve derin ağlardaki ilgili bilgi akışının modülasyonunu sağlanması GRU'daki başlıca zorluklardır. Bu çalışmada, ilgili bilgilerin çoklu GRU katmanları kullanılarak aktarılmasını sağlamak, ve kaybolan gradyan sorununun üstesinden gelmek için resnet tabanlı bir derin GRU yaklaşımı önerilmektedir. Derin GRU'nun ardışık katmanları arasında artık bağlantılar kullanılmasıyla alt katmanlardan üst katmanlara doğru gradyan akışının iyileştirilmesi sağlanmıştır. Yaygın olarak kullanılan MSCOCO veri seti üzerindeki deneysel araştırmalar, önerilen yaklaşımın son yaklaşımlarla karşılaştırılabilir performans sağladığını göstermiştir. Ayrıca bu yaklaşım, internet bağlantısı olmaksızın altyazı oluşturma olanağı sunan ve sesle kontrol edilebilen bir arayüzü olan kendi tasarladığımız Android uygulamamıza *CaptionEye* gömülmüştür.
Anahtar Kelimeler: Kapılı Tekrarlayan Birim, Artık Bağlantı, Görüntü Altyazılama, Android Uygulama.

1. Introduction

Image captioning focuses on expressing images with linguistically and semantically proper sentences (Fetiler, Çaylı, Moral, Kılıç, & Aytuğ, 2021; Keskin, Çaylı, Moral, Kılıç, & Aytuğ, 2021), which has found applications in visual question answering (Anderson et al., 2018), image indexing (Chang, 1995), and virtual assistants (Aydın, Çaylı, Kılıç, & Aytuğ Onan, 2022; Baran, Moral, & Kılıç, 2021; Çaylı, Makav, Kılıç, & Onan, 2020; Keskin, Moral, Kılıç, & Onan, 2021; Makav & Kılıç, 2019b).

Recent studies mainly benefit from retrieval-based, template-based, and neural encoder-decoder frameworks for image captioning. The retrieval-based framework creates a candidate caption set from reference captions in the dataset similar to the input image. A caption that describes the most semantic information of the input image is selected from the candidate set (Yang et al., 2020). The dependency of the candidate set on the reference captions results in meaningless captions for images different from those in the training set.

The template-based framework predicts tags utilizing subjects, objects, and verbs from the image and then employs predefined language templates to generate a caption. However, the utilization of the framework causes no diversity in the generated captions due to the limited number of templates (Yu, Li, Yu, Huang, & technology, 2019). The neural encoder-decoder framework, which conventionally consists of a convolutional neural network (CNN) and recurrent neural network (RNN), overcomes the issues in retrieval-based and template-based frameworks (Makav & Kılıç, 2019a; Mao et al., 2014) because it conveys visual information of images as a latent vector for effective caption generation. In general, many researchers exploit CNN architectures pre-trained on largescale image classification datasets such as Inception-v3 (Szegedy, Vanhoucke, Ioffe, Shlens, & Wojna, 2016), Xception (Chollet, 2017), and NASNet (Qin & Wang, 2019) in the encoder because training from scratch is computationally expensive and time-consuming.

The latent vector is fed to the RNN-based decoder utilizing injection techniques for caption generation. There are four injection techniques for image captioning: init-inject, pre-inject, par-inject, and merge (Tanti, Gatt, & Camilleri, 2018). For the init-inject, the latent vector which contains the image features is fed to the initial state of the RNN (Devlin et al., 2015; Liu, Zhu, Ye, Guadarrama, & Murphy, 2017). The pre-inject employs the latent vector as input to the RNN priorly to the words (Nina & Rodriguez, 2015; Rennie, Marcheret, Mroueh, Ross, & Goel, 2017; Vinyals, Toshev, Bengio, & Erhan, 2015). In the par-inject, the latent vector is concatenated with a word embedding layer output and fed to the RNN as input (Donahue et al., 2015; Yao, Pan, Li, Qiu, & Mei, 2017). The merge adds the latent vector to linguistic features coming from the RNN before the output layer of the decoder (Baran et al., 2021), (Mao et al., 2014), (Mao et al., 2015). RNN is specifically developed to deal with long sequence data. However, as the sequence gets longer, vanilla RNN encounters issues of vanishing and exploding gradients (Bengio, Simard, & Frasconi, 1994; Q. Wang, Bu, & He, 2020). Therefore, gated RNNs such as Long-short-term memory (LSTM) (Hochreiter & Schmidhuber, 1997) and GRU (Chung, Gulcehre, Cho, & Bengio, 2014) have been developed to overcome these issues. LSTM provides high performance in language modeling studies, such as machine translation and captioning, with input, output, and forget gates. GRU combines the forget and input gates into a single update gate, making the GRU utilize fewer parameters than the LSTM. Therefore, GRU-based networks prominently converge on large-scale datasets faster than LSTM-based (Kılıç, 2021; B. Wang, Kong, Guan, & Xiong, 2019). Stacking RNN layers improves the network to capture more relevant features, leading to a meaningful caption (Rahman, Srikumar, & Smith, 2018; Sagheer & Kotb, 2019). Despite the promising results with multiple RNN layers, retaining the gradient flow becomes challenging due to increased number of layers.

In this study, we propose a resnet-based deep gated recurrent unit approach under the neural encoder-decoder framework for image captioning. The approach employs residual connections between subsequent layers on the decoder side to maintain gradient flow. Inception-v3 CNN architecture pre-trained on the ImageNet dataset is utilized on the encoder side. We used the MSCOCO Captions dataset (Lin et al., 2014) for experiments and evaluated the efficacy with performance metrics such as CIDEr, SPICE, METEOR, ROUGE-L, and BLEU-n (1, 2, 3, 4).

The rest of the paper is organized as follows: Section 2 presents the proposed image captioning approach with theoretical foundations and our custom-designed Android application CaptionEye, which generates a caption without an internet connection. Section 3 introduces the setup of the dataset, performance metrics, and results. Closing remarks are given in Section 4.

2. Methodology

In this section, the proposed image captioning approach under the encoder-decoder framework is first introduced. Then, we present the Android application *CaptionEye* running the proposed approach without an internet connection.

2.1. The Proposed Approach

We utilize the Inception-v3 CNN architecture to extract the features for a given image in the encoder, and then an RNN based decoder generates a caption in the proposed approach.

The Inception-v3 architecture consists of convolution, pooling, and linear layers and has an input size of 3-by-299-by-299. Furthermore, the output of the global average pooling layer is taken as the image feature, which has a size of 2048.

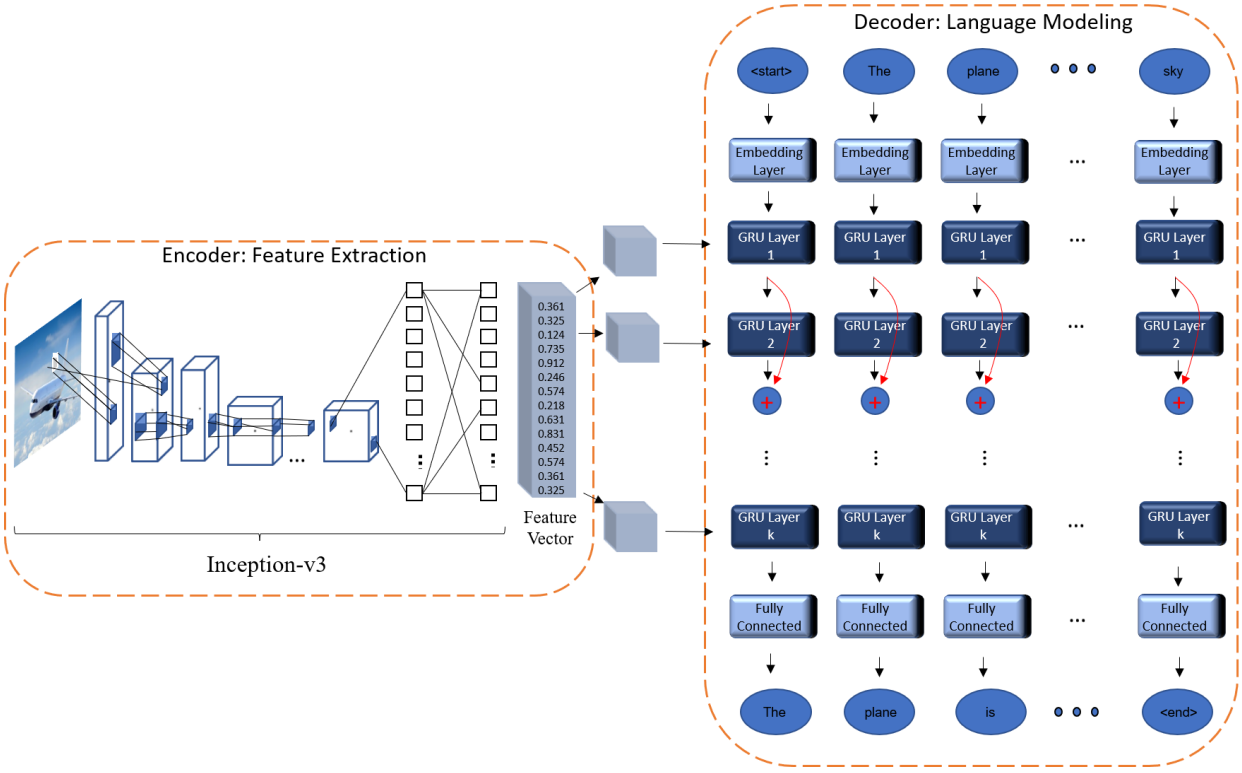

Figure 1. The Proposed Image Captioning Approach

Then, the feature is passed into the decoder as the latent vector. On the other hand, the decoder consists of embedding, multi-layer GRU, and linear layers. The embedding layer aims to represent the words as meaningful vectors, which founds effective use in language modeling. A GRU is a gated RNN with reset r_t, update z_t, and new u_t gates. The GRU updates the hidden state as shown in Equation 1.

$$\begin{aligned} r_t &= \sigma(W_r x_t + W_r h_{t-1}) \\ z_t &= \sigma(W_z x_t + W_z h_{t-1}) \\ h_t &= (1 - z_t) h_{t-1} + z_t u_t \\ u_t &= tanh\big(W_u x_t + W_u (r_t h_{t-1})\big) \end{aligned} \quad (1)$$

where x_t is the input, and h_t is the hidden state at time t. Similarly, W_r, W_z, and W_u are the weights for the reset, update, and new gates, respectively. Furthermore, tanh and sigmoid are the activation functions denoted as $tanh$, and σ, respectively.

$$\begin{aligned} x_t^l, h_t^l &= GRU_l(x_t^{l-1},\ h_{t-1}^l) \\ x_t^{l+1}, h_t^{l+1} &= GRU_{l+1}(x_t^l + x_t^{l-1}, h_{t-1}^{l+1}) \end{aligned} \quad (2)$$

where superscript l represents the order of the GRU layer. Deep GRU consists of multiple GRU layers stacked on top of each other. Residual connections are employed between the layers as shown in Equation (*2*).

The linear layer consists of weight and bias and outputs a probability distribution over words. $\mathbf{Y} = y_1, y_2, y_{3,\ldots\ldots}, y_n$ where y_n corresponds to the n-th word which represents a ground-truth sentence. Similarly, $\hat{\mathbf{Y}}$ is the sequential prediction of the approach. The cross entropy loss, a combination of softmax activation function and negative log-likelihood loss($NLLL$), is utilized in training. Therefore, the loss is calculated as $loss$=$NLLL$($softmax(Y)$, $softmax(\hat{Y})$). In addition, the stochastic gradient descent algorithm is utilized for training.

2.2. Android-based Application

The developed Android application called *CaptionEye* offers to generate captions on smartphones without an internet connection, leading to an improved response time. First, the approach is quantized to accelerate the inference time. Then the approach is converted to a script to infer in the application. *CaptionEye* can take images using the camera or gallery. In addition, it supports various languages with speech command recognition. Screenshots of the *CaptionEye* are given in Figure 2.

3. Experimental Evaluations

This section presents performance evaluations of the proposed approach using the MSCOCO Captions dataset.

3.1. Dataset and Performance Metrics

Here, we evaluate the proposed approach on the MSCOCO Captions as it a large-scale dataset that contains 118287 training and 5000 validation images, and each has at least five corresponding reference captions. We choose six to fifteen word captions in the training set to ensure consistency in the generated caption lengths.

Decoder Design	# of Decoder Layer	CIDEr	BLEU-4	BLEU-3	BLEU-2	BLEU-1	ROUGE-L	METEOR	SPICE
Multi-layer GRU with residual connections	3-Layer	0.692	0.210	0.310	0.455	0.648	0.471	0.213	0.144
	4-Layer	0.672	0.209	0.308	0.454	0.648	0.471	0.210	0.140
	5-Layer	0.680	0.208	0.309	0.455	0.651	0.470	0.209	0.142
	6-Layer	0.679	0.210	0.310	0.454	0.647	0.471	0.210	0.140
	7-Layer	0.680	0.210	0.309	0.454	0.646	0.469	0.210	0.139
	8-Layer	0.722	**0.223**	**0.325**	0.471	0.661	**0.481**	**0.215**	**0.146**
	9-Layer	0.718	0.216	0.316	0.463	0.656	0.475	0.214	0.145
	10-Layer	**0.723**	0.221	**0.325**	**0.472**	**0.665**	0.479	0.213	**0.146**
Multi-layer GRU without residual connections	3-Layer	0.677	0.211	0.311	0.453	0.645	0.471	0.210	0.139
	4-Layer	0.676	0.208	0.306	0.450	0.644	0.470	0.211	0.140
	5-Layer	**0.716**	0.224	0.326	0.471	0.660	0.480	**0.215**	**0.145**
	6-Layer	0.674	0.211	0.311	0.458	0.652	0.470	0.209	0.138
	7-Layer	0.673	0.211	0.314	0.462	0.657	0.475	0.208	0.140
	8-Layer	0.705	**0.228**	**0.332**	**0.479**	**0.669**	**0.483**	0.213	0.143
	9-Layer	0.681	0.213	0.315	0.460	0.650	0.474	0.210	0.140
	10-Layer	0.702	0.222	0.323	0.468	0.659	0.480	0.213	0.142

Table 1. Experimental evaluations on residual connection

BLEU-n (n = 1, 2, 3, 4), ROUGE-L, SPICE, METEOR, and CIDEr performance metrics are employed to check the accuracy of the proposed approach. BLEU-n measures the n-gram overlap between generated and reference captions developed initially for machine translation. Similarly, METEOR is developed for machine translation and utilizes the harmonic average of unigram matches between precision and recalls. ROUGE-L is a text summary performance metric that measures the longest common subsequence between a generated caption and references. SPICE is a semantic performance metric that first parses each reference sentence and then evaluates the objects, attributes, and relationships in the generated captions, rather than directly comparing a generated caption with a set of reference sentences for syntactic compatibility. CIDEr evaluates the consensus between a caption and references, exploiting sentence similarity to capture grammatical accuracy and saliency concepts. We choose the CIDEr to compare the main findings as it is the default metric in the MSCOCO Captions dataset evaluations.

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-	ROUGE-L	SPICE	METEOR	CIDEr
(Chen et al., 2018)	0.505	0.308	0.191	0.121	-	-	-	0.600
(You, Jin, & Luo, 2018)	0.511	0.322	0.207	0.136	0.390	-	0.170	0.655
Proposed Approach	**0.655**	**0.472**	**0.325**	**0.221**	**0.479**	**0.146**	**0.213**	**0.723**

Table 2. Comparison with state-of-the-art

3.2. Result and Discussion

We evaluate the effect of the residual connected deep GRU with performance metrics and qualitative analysis on the validation set of the dataset. The deep GRU is utilized from three to ten layers with and without residual connections in the experiments.

In Table 1, we present the scores of the proposed approach across multiple performance metrics where "with residual connections" achieves the highest CIDEr score in the 10-layer. However, the approach "without residual connections" reaches its finest in the 5-layer. This discrepancy in the scores could be attributed to the benefits of residual connections.

In Table 2, we report a comparison of the proposed residual connected multi-layer GRU with encoder-decoder approaches. Our proposed approach outperforms (Chen et al., 2018) and (You et al., 2018) in terms of all performance metrics.

Compared with the "without residual connections", the "with residual connections" achieves more accurate caption generation. For the first image of Table 3, the proposed approach generates the caption "a city street with a lot of people walking around it" which is a meaningful description. On the other hand, the caption generated by the "without residual connections" refers to "a large clock tower" which is defective information for the image. Furthermore, "with residual connections" expresses the image as "a kite is flying in the sky" where the phrase "a kite" in the caption causes misinformation because there are two kites in the image. However, "without residual connections" causes wrong knowledge with the phrase "a person standing in a field" by recognizing unfound objects. The results verify that the proposed approach emerges more semantical objects and attributes in the captions.

Figure 2 presents screenshots of *CaptionEye*. The home page of the application as shown in Figure 2 (a). The microphone icon on the main screen leads the user to the speech command recognition, as shown in Figure 2 (b). The gallery icon on the main screen allows the user to access the phone gallery shown in Figure 2 (c), while Figure 2 (d) shows the caption generated by the application. The speaker icon on the main screen narrates the generated caption. The three dots icon on the right of the main screen opens the settings window in Figure 2 (e): language selection, application background selection, voice speed, and pitch adjustments. Figure 2 (f) shows the language selection screen. Twenty-one different language options are offered in the application. Figure 2 (g) shows the image capturing screen, which opens with the camera icon on the main page. Figure 2 (h) shows the generated caption with the selected language option.

	
Reference Captions: (1) A chinese street with a mcdonalds in the back drop. (2) People and buses on a city street under cloudy skies. (3) A large open area with concrete floor and a mcdonalds in the background with chinese writing on the building. (4) An asian city square, with people, buses, and a mcdonald's. (5) A red bus driving down a busy city street surrounded by tall buildings.	(1) Two kites, one with a smily face fly high in the sunny sky. (2) Two kites flying directly overhead against a clear blue sky. (3) Many kites can be seen in the air through umbrellas. (4) Two kites flying in the sky over an open and closed umbrella. (5) A few kites flying in the blue sky.
Generated Captions: **Residual connections:** a city street with a lot of people walking around it. **Without residual connections:** a crowded city street with a large clock tower.	**Residual connections:** a kite is flying in the sky. **Without residual connections:** a person standing in a field flying a colorful kite.

Table 3. Samples with generated and reference captions

4. Conclusion

In this paper, a residual-connected multi-layer GRU under the encoder-decoder framework for image captioning, has been proposed. The proposed approach includes an Inception-v3 as an encoder for feature extraction of the image and a residual connected GRU-based decoder to generate the corresponding caption. Residual connections carry the features in the multi-layer GRU from lower to upper layers inducing a more valid gradient flow. We evaluated the approach on the MSCOCO Captions dataset, which demonstrated that residual connections performance improved significantly and achieved state-of-the-art performance. Additionally, we developed an Android application named *CaptionEye* utilizes the proposed approach with a user-friendly interface that has great potential for the visually impaired in daily activities.

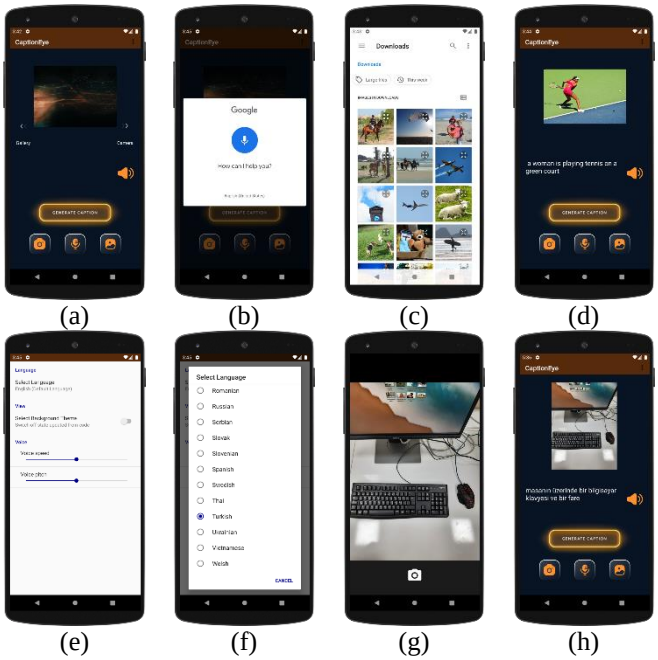

Figure 2. In (a) home screen, in (b) voice command screen, in (c) gallery, in (d) main screen with English caption, in (e) settings screen, in (f) language options, in (g) camera, in (h) main screen with image captions in the selected language are shown.

5. Acknowledge

This research was supported by the Scientific and Technological Research Council of Turkey (TUBITAK)-British Council (The Newton-Katip Celebi Fund Institutional Links, Turkey-UK project: 120N995) and TUBITAK 2209-B Industry Oriented Research Project Support Programme for Undergraduate Students with project no: 1139B412100443.

References

Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., & Zhang, L. (2018). *Bottom-up and top-down attention for image captioning and visual question answering.* Paper presented at the Proceedings of the IEEE conference on computer vision and pattern recognition.

Aydın, S., Çaylı, Ö., Kılıç, V., & Aytuğ Onan. (2022). Sequence-to-sequence video captioning with residual connected gated recurrent units. *European Journal of Science and Technology*((35), 380–386.

Baran, M., Moral, Ö. T., & Kılıç, V. (2021). Akıllı Telefonlar için Birleştirme Modeli Tabanlı Görüntü Altyazılama. *European Journal of Science and Technology*(26), 191-196.

Bengio, Y., Simard, P., & Frasconi, P. J. I. t. o. n. n. (1994). Learning long-term dependencies with gradient descent is difficult. *5*(2), 157-166.

Chang, S.-F. (1995). *Compressed-domain techniques for image/video indexing and manipulation.* Paper presented at the Proceedings., International Conference on Image Processing.

Chen, T., Zhang, Z., You, Q., Fang, C., Wang, Z., Jin, H., & Luo, J. (2018). ``*Factual''or``Emotional'': Stylized Image*

Captioning with Adaptive Learning and Attention. Paper presented at the Proceedings of the European Conference on Computer Vision (ECCV).

Chollet, F. (2017). *Xception: Deep learning with depthwise separable convolutions.* Paper presented at the Proceedings of the IEEE conference on computer vision and pattern recognition.

Chung, J., Gulcehre, C., Cho, K., & Bengio, Y. J. a. p. a. (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling.

Çaylı, Ö., Makav, B., Kılıç, V., & Onan, A. (2020). *Mobile Application Based Automatic Caption Generation for Visually Impaired.* Paper presented at the International Conference on Intelligent and Fuzzy Systems.

Devlin, J., Cheng, H., Fang, H., Gupta, S., Deng, L., He, X., . . . Mitchell, M. J. a. p. a. (2015). Language models for image captioning: The quirks and what works.

Donahue, J., Anne Hendricks, L., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., & Darrell, T. (2015). *Long-term recurrent convolutional networks for visual recognition and description.* Paper presented at the Proceedings of the IEEE conference on computer vision and pattern recognition.

Fetiler, B., Çaylı, Ö., Moral, Ö. T., Kılıç, V., & Aytuğ, O. (2021). Video Captioning Based on Multi-layer Gated Recurrent Unit for Smartphones. *European Journal of Science and Technology*(32), 221-226.

Hochreiter, S., & Schmidhuber, J. J. N. c. (1997). Long short-term memory. *9*(8), 1735-1780.

Keskin, R., Çaylı, Ö., Moral, Ö. T., Kılıç, V., & Aytuğ, O. (2021). A Benchmark for Feature-injection Architectures in Image Captioning. *European Journal of Science and Technology*(31), 461-468.

Keskin, R., Moral, Ö. T., Kılıç, V., & Onan, A. (2021). *Multi-GRU Based Automated Image Captioning for Smartphones.* Paper presented at the 2021 29th Signal Processing and Communications Applications Conference (SIU).

Kılıç, V. (2021). Deep Gated Recurrent Unit for Smartphone-Based Image Captioning. *Sakarya University Journal of Computer Information Sciences, 4*(2), 181-191.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., . . . Zitnick, C. L. (2014). *Microsoft coco: Common objects in context.* Paper presented at the European conference on computer vision.

Liu, S., Zhu, Z., Ye, N., Guadarrama, S., & Murphy, K. (2017). *Improved image captioning via policy gradient optimization of spider.* Paper presented at the Proceedings of the IEEE international conference on computer vision.

Makav, B., & Kılıç, V. (2019a). *A new image captioning approach for visually impaired people.* Paper presented at the 2019 11th International Conference on Electrical and Electronics Engineering (ELECO).

Makav, B., & Kılıç, V. (2019b). *Smartphone-based image captioning for visually and hearing impaired.* Paper presented at the 2019 11th International Conference on Electrical and Electronics Engineering (ELECO).

Mao, J., Wei, X., Yang, Y., Wang, J., Huang, Z., & Yuille, A. L. (2015). *Learning like a child: Fast novel visual concept learning from sentence descriptions of images.* Paper presented at the Proceedings of the IEEE international conference on computer vision.

Mao, J., Xu, W., Yang, Y., Wang, J., Huang, Z., & Yuille, A. J. a. p. a. (2014). Deep captioning with multimodal recurrent neural networks (m-rnn).

Nina, O., & Rodriguez, A. (2015). *Simplified LSTM unit and search space probability exploration for image description.* Paper presented at the 2015 10th International Conference on Information, Communications and Signal Processing (ICICS).

Qin, X., & Wang, Z. J. a. p. a. (2019). Nasnet: A neuron attention stage-by-stage net for single image deraining.

Rahman, A., Srikumar, V., & Smith, A. D. J. A. e. (2018). Predicting electricity consumption for commercial and residential buildings using deep recurrent neural networks. *212*, 372-385.

Rennie, S. J., Marcheret, E., Mroueh, Y., Ross, J., & Goel, V. (2017). *Self-critical sequence training for image captioning.* Paper presented at the Proceedings of the IEEE conference on computer vision and pattern recognition.

Sagheer, A., & Kotb, M. J. N. (2019). Time series forecasting of petroleum production using deep LSTM recurrent networks. *323*, 203-213.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). *Rethinking the inception architecture for computer vision.* Paper presented at the Proceedings of the IEEE conference on computer vision and pattern recognition.

Tanti, M., Gatt, A., & Camilleri, K. P. J. N. L. E. (2018). Where to put the image in an image caption generator. *24*(3), 467-489.

Vinyals, O., Toshev, A., Bengio, S., & Erhan, D. (2015). *Show and tell: A neural image caption generator.* Paper presented at the Proceedings of the IEEE conference on computer vision and pattern recognition.

Wang, B., Kong, W., Guan, H., & Xiong, N. N. J. I. A. (2019). Air quality forecasting based on gated recurrent long short term memory model in Internet of Things. *7*, 69524-69534.

Wang, Q., Bu, S., & He, Z. J. I. T. o. I. I. (2020). Achieving predictive and proactive maintenance for high-speed railway power equipment with LSTM-RNN. *16*(10), 6509-6517.

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., . . . Macherey, K. J. a. p. a. (2016). Google's neural machine translation system: Bridging the gap between human and machine translation.

Yang, M., Liu, J., Shen, Y., Zhao, Z., Chen, X., Wu, Q., & Li, C. J. I. T. o. I. P. (2020). An ensemble of generation-and retrieval-based image captioning with dual generator generative adversarial network. *29*, 9627-9640.

Yao, T., Pan, Y., Li, Y., Qiu, Z., & Mei, T. (2017). *Boosting image captioning with attributes.* Paper presented at the Proceedings of the IEEE international conference on computer vision.

You, Q., Jin, H., & Luo, J. J. a. p. a. (2018). Image captioning at will: A versatile scheme for effectively injecting sentiments into image descriptions.

Yu, J., Li, J., Yu, Z., Huang, Q. J. I. t. o. c., & technology, s. f. v. (2019). Multimodal transformer with multi-view visual representation for image captioning. *30*(12), 4467-4480.