Automated Image Captioning with Multi-layer Gated Recurrent Unit

Özge Taylan Moral[1], Volkan Kılıç[1*], Aytuğ Onan[2], Wenwu Wang[3‡]
[1]Electrical and Electronics Engineering Graduate Program, Izmir Katip Celebi University, Turkey
[2]Department of Computer Engineering, Izmir Katip Celebi University, Turkey
[3]Centre for Vision, Speech and Signal Processing (CVSSP), University of Surrey, UK
Email: *volkan.kilic@ikcu.edu.tr; ‡w.wang@surrey.ac.uk

***Abstract*—Describing the semantic content of an image via natural language, known as image captioning, has recently attracted substantial interest in computer vision and language processing communities. Current image captioning approaches are mainly based on an encoder-decoder framework in which visual information is extracted by an image encoder and captions are generated by a text decoder, using convolution neural networks (CNN) and recurrent neural networks (RNN), respectively. Although this framework is promising for image captioning, it has limitations in utilizing the encoded visual information for generating grammatically and semantically correct captions in the RNN decoder. More specifically, the RNN decoder is ineffective in using the contextual information from the encoded data due to its limited ability in capturing long-term complex dependencies. Inspired by the advantage of gated recurrent unit (GRU), in this paper, we propose an extension of conventional RNN by introducing a multi-layer GRU that modulates the most relevant information inside the unit to enhance the semantic coherence of captions. Experimental results on the MSCOCO dataset show the superiority of our proposed approach over the state-of-the-art approaches in several performance metrics.**

***Index Terms*—convolutional neural network, gated recurrent unit, image captioning, recurrent neural network**

I. Introduction

Automated generation of grammatically correct and meaningful descriptions of images using computer vision and natural language processing is known as image captioning, which has recently received much attention due to its potential applications, such as visual search, virtual reality, augmented reality, image retrieval and indexing [1]–[5].

Earlier studies on image captioning mainly use template-based, retrieval-based and encoder-decoder framework based approaches [6]. In template-based approaches, the visual information of an image is extracted via the detection of objects, scenes and attributes, which is then used to generate a fixed-length sentence from the predefined templates [7]. However, the quality of the generated sentences depends on the object detectors and templates. This limitation is addressed in the retrieval-based methods, where reference captions of similar images are retrieved from the retrieval library using the visual information of the image, for generating variable-length and semantic captions [8]. The generated caption using this method, however, is limited by the similarities between the input image and the images from the retrieval library, as a result, the descriptions generated for the input image may still be far from being consistent with its contents.

Mao et al. [9] proposed an effective encoder-decoder method where CNN and RNN are combined to address the limitations of the template-based and retrieval-based approaches. In this method, two sub-networks (i.e. encoder and decoder networks) are used, where feature representation of an image is extracted in the encoder using a CNN, while an RNN is used in the decoder to generate captions from this representation. The CNN architectures have a high capacity of learning from a large number of images, and are currently the architecture popularly chosen for the encoder design [10], [11]. ResNet152 v2 [12], Xception [13], NASNet-Large [14] and Inception-v3 [15], [16] are the popular architectures chosen for extracting the image features as a vector from the second to last fully connected layer. In the RNN-based decoders, image features are utilized as visual information to generate natural language captions word-by-word. As conventional RNNs have gradient vanishing and exploding problems, they cannot capture long-term dependencies. Long short-term memory (LSTM) and GRU are the RNN architectures that solve these problems with the gating mechanism. A hierarchical LSTM architecture has been introduced to describe the salient objects in an image using phrases [17]. Another work proposed an extension of the LSTM by adding the extracted semantic information of the image as extra input to each unit of the LSTM block [18]. In addition to the RNN-based decoder, a CNN-based decoder is proposed to integrate a vision module using VGG16 and an attention module to connect CNNs for caption generation [19].

In the caption generation process, the visual and linguistic information can be fed into the RNN-based decoder with different feature-injection architectures, such as pre-inject, par-inject, merge, and init-inject [20]. The visual information is used as the initial hidden state of the RNN for init-inject [21], [22], while it is fed to the first input of the RNN in the pre-inject [23]–[25]. The visual information with the linguistic information is used in parallel as an input to the RNN for par-inject [26], [27]. Different from these architectures, the merge architecture takes the visual information after the RNN processes [9], [28], [29]. It is reported that init-inject offers

Identify applicable funding agency here. If none, delete this.

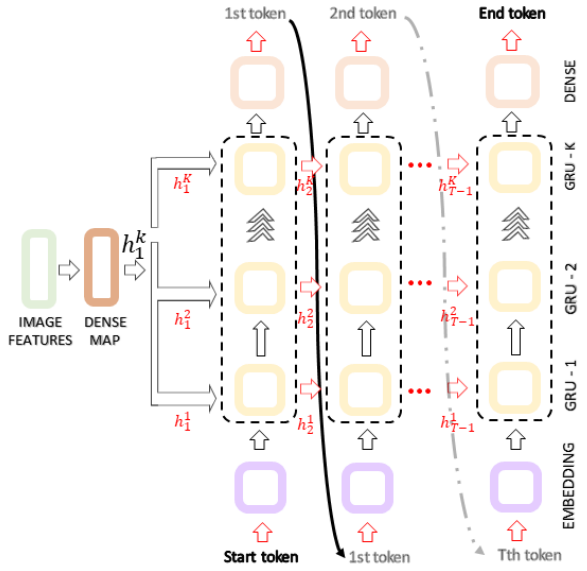

Fig. 1. The proposed multi-layer GRU based decoder

better performance compared to the other architectures in terms of generation and retrieval measures [20].

In [15], the Inception-v3 CNN is used to extract visual information, and the GRU based RNN decoder is designed under init-inject architecture, where image captions are generated with additional annotations in training. To retain more semantic information of the image without additional annotations, in this paper, we propose a new encoder-decoder image captioning approach with Inception-v3 and multi-layer GRU under init-inject architecture. The motivation behind the multi-layer structure is to modulate the most relevant information flow within GRU. The increment of layers within the multi-layer GRU provides representations of increased complexity in learning sequential data, and thus giving an enhanced prediction model [30], [31]. We employ different layer-size of GRU to investigate the optimal number for image captioning.

The remainder of the paper is organized as follows. Section II introduces the proposed multi-layer GRU for image captioning. Section III presents the dataset, performance metrics and experimental results. Concluding remarks with future works are given in Section IV.

II. Proposed Multi-Layer GRU for Image Captioning

The proposed image captioning approach follows the encoder-decoder architecture by combining an image encoder with a language decoder. The image encoder employs the convolutional and pooling layers of the CNN architectures to extract the feature representation of an input image. CNN architectures such as Inception-v3, NASNet-Large, Xception, and ResNet152 v2 are commonly used in the encoder design. Inception-v3 is a 42-layered deep CNN architecture that uses the asymmetric approach to decompose a kernel of large-scale convolution into a small-scale kernel of convolution [16]. The Inception-v3 model is utilized here as an image feature extractor that returns a 2048-element vector after the average pooling layer. Then, these features of each input image will be fed into the decoder to generate a caption word-by-word.

Our proposed multi-layer GRU based decoder has three main parts as embedding layer, GRUs, and dense layer, illustrated in Fig. 1. The init-inject architecture is applied to design the proposed decoder. Image features and linguistic features are fed from the dense map and embedding layer, respectively. The multi-layer GRU combines K GRUs for $k = 1, ..., K$, while h_t^k is defined as the hidden vector for the kth GRU layer and the variable-length $t = 1, ..., T$. The dense map reduces the 2048-element vector to 512 for feeding the multi-layer GRU at $t = 1$ for each initial hidden state. The updated hidden state feeds the multi-layer GRU from the previous iteration to the next iteration. The linguistic features are obtained as a meaningful embedding vector by the embedding layer using the tokens. At $t = 1$, the embedding layer utilizes the start token to lead the input of the first GRU layer. The output vector of the first GRU layer is fed to the next layer. This process is repeated K-times. The output of the multi-layer GRU is fed to the dense layer to compute the prediction probabilities and generate the following token for the caption. The dense layer is used to generate the first token to be utilized in the next step, and the generation process continues T-times to reach the end token. All generated tokens are matched with corresponding words from the vocabulary, created in pre-processing steps before the training.

III. Experiments

This section presents the evaluation results of the proposed captioning approach on the MSCOCO dataset [32], and the performance comparison with state-of-the-art approaches.

A. Dataset and Performance Metrics

The studies on captioning require large image datasets such as Flickr [33], VizWiz-Captions [34] and MSCOCO [32], for the performance evaluation. Flickr8k and Flickr30k contain 8000 and 31783 images, whereas VizWiz-Captions consists of 39181 images captured by blind people, paired with five reference captions. The MSCOCO dataset is an extensive dataset containing 118287 images for training, 41000 images for testing and 5000 images for validation [32], each having five reference captions. MSCOCO is employed to evaluate our proposed image captioning approach due to its high number of images with diverse contents.

BLEU-n [35], CIDEr [36], METEOR [37], ROUGE-L [38] and SPICE [39] are the well-known metrics used to analyze the performance of captioning approaches. BLEU-n (n = 1, ..., 4) and METEOR are machine translation metrics where BLEU-n uses n-gram (e.g. BLEU-2 for 2-grams) pairs to compare a machine-generated caption with the human-generated ground truth captions [35] while METEOR generalizes unigram matches between a machine-generated caption and a human-generated ground truth captions [37]. ROUGE-L uses the longest common subsequence to measure sentence-to-sentence similarity between the generated caption and a set of reference

TABLE I
COMPARISON OF INITIAL AND FINE-TUNING PARAMETERS

CNN	# of layers	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-L	METEOR	SPICE	CIDEr
Inception-v3 with the initial parameters	3	0.687	0.502	0.356	0.252	**0.499**	0.226	0.156	0.821
	6	**0.688**	**0.503**	**0.357**	**0.253**	0.499	**0.227**	**0.158**	**0.821**
	9	0.528	0.314	0.176	0.099	0.412	0.149	0.086	0.426
	12	0.532	0.321	0.183	0.107	0.415	0.151	0.087	0.432
	15	0.529	0.319	0.182	0.104	0.415	0.149	0.087	0.429
Inception-v3 with the fine-tuning parameters	3	0.694	0.512	0.364	0.258	0.498	0.226	0.157	0.839
	6	0.687	0.506	0.364	0.261	0.497	0.225	0.155	0.824
	9	**0.700**	**0.521**	**0.376**	**0.271**	**0.505**	**0.230**	**0.160**	**0.843**
	12	0.537	0.327	0.186	0.107	0.412	0.153	0.087	0.454
	15	0.552	0.337	0.191	0.110	0.419	0.155	0.092	0.458

TABLE II
COMPARISON OF OUR PROPOSED APPROACH WITH SOME STATE-OF-THE-ART APPROACHES ON THE MSCOCO DATASET

	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-L	METEOR	SPICE	CIDEr
[31]	0.652	0.470	0.330	0.232	0.484	-	-	0.775
[18]	0.663	0.485	0.354	0.262	-	0.230	-	0.813
[17]	0.666	0.489	0.355	0.258	0.497	0.231	**0.165**	0.821
[19]	0.688	0.513	0.370	0.265	**0.507**	**0.234**	-	0.839
Proposed	**0.700**	**0.521**	**0.376**	**0.271**	0.505	0.230	0.160	**0.843**

captions [38]. CIDEr and SPICE are designed especially for image captioning tasks. CIDEr ensures the consistency of a generated caption, calculating the different weights of n-gram words with term-frequency-inverse document frequency [36], whereas SPICE measures the semantic correctness of the caption using scene-graphs that contain objects, attributes, and relationships between them [39]. In the evaluation of image captioning, CIDEr is found to be more appropriate to measure the correlation between generated and reference captions [36]. In this paper, therefore, the results are sorted based on CIDEr metric.

B. Results & Discussion

The proposed multi-layer GRU based decoder under init-inject architecture has been first analyzed with the Inception-v3 encoder in order to find the optimum layer size. The Inception-v3 with five different layer-sized GRU were evaluated in terms of BLEU-n, CIDEr, METEOR, SPICE, and ROUGE-L metrics for configurations based on the selection of best embedding vector and vocabulary size.

Our proposed multi-layer based decoder utilizes linguistic features for image caption generation. The vocabulary and embedding vector size are two critical parameters related to linguistic information affecting captioning performance. The embedding vector is typically set to 50 and 300 dimensions and the small-dimensional vector does not capture the word relations completely, whereas the large-dimensional vector causes the overfitting [40]. The vocabulary size is determined based on the number of common words in all reference captions and usually varies from 10000 to 40000 words [41]. First, the embedding vector size and the vocabulary size are chosen as 100 and 750, respectively. The Inception-v3 with five different layer-sized GRU is employed for fixed parameters and evaluated under performance metrics given in the first row of Table I. To fine-tune this performance, our proposed multi-layer GRU based decoder is tested under ten different vocabulary sizes, including 250, 500, 750, 1000, 2000, 3000, 5000, 10000, 20000, and 40000, and eight different embedding vector sizes, namely, 25, 50, 75, 100, 125, 150, 200, and 250.

For optimization, the Inception-v3 with 3-layer GRU based decoder is used as an initial configuration in the decoder. First, this initial configuration was evaluated under different performance metrics with ten different vocabularies and the embedding vector of fixed-size, as 100. The best CIDEr metric was observed when the vocabulary size was 20000. Then, the initial configuration was evaluated under the same performance metrics with eight different embedding vectors and the vocabulary of fixed-size, as 20000. The best CIDEr metric was observed when the embedding vector size was 100. Applying the same procedure to the 6, 9, 12, and 15 layer GRU, the optimum parameters were determined as 100, 250, 125, and 50 for the embedding vector size; 2000, 20000, 2000, and 40000 for the vocabulary size, respectively.

The fine-tuning performance listed in the second row of Table I indicates that the CIDEr metric gradually increases up to the 9-layer GRU, where the maximum level has been reached. Among all the configurations, 9-layer GRU outperforms the other layer-sized GRU in terms of all metric scores. However, the captioning performance is degraded for the number of layers of 12 and 15, causing the test to stop with more layers,

TABLE III
SAMPLES IMAGES FROM MSCOCO WITH REFERENCE AND GENERATED CAPTIONS

Reference Captions			
(1) A horse standing on top of a lush green field	**(1)** a herd of zebras crossing a shallow part of a river	**(1)** Several colorful kites in the sky by several persons in the ground	**(1)** A couple of snow skiers are casting a shadow on the snow
(2) a horse grazing in a green pasture bordered by mountains	**(2)** A herd of zebra crossing a river with trees in the background	**(2)** Kites flying over a busy beach area on clear day	**(2)** Two people in ski gear standing at the top of a mountain
(3) A horse grazing with mountains behind him	**(3)** A herd of zebra crossing a river near a forest	**(3)** A group of people standing on top of a sandy beach	**(3)** Two people wearing skis on a snowy slope
(4) A horse is grazing in a grassy field with a view of mountains	**(4)** A group of zebras are crossing a stream	**(4)** a large gathering of people flying their kites	**(4)** Two people with skis on at the top of a mountain
(5) A horse in a grassy field set against a foggy mountain range	**(5)** Some black and white zebras crossing a shallow stream	**(5)** People on a beach beneath many colorful kites	**(5)** There are people standing in the snow on skis
Generated Captions			
3-layer a horse grazing on a lush green field	**3-layer** a herd of zebra drinking from a river	**3-layer** a crowd of people standing on top of a sandy flying	**3-layer** two people on skis standing on a snowy hill
6-layer a horse is grazing in a field with mountains in and and and and and and and and and	**6-layer** a herd of zebra standing next to a river	**6-layer** a group of people standing on top of a beach beach	**6-layer** a man and a woman are standing on skis
9-layer a horse standing in a field with mountains in the background	**9-layer** a herd of zebra are crossing a river	**9-layer** a group of people standing on a beach with kites	**9-layer** two skiers posing for a picture on a mountain
12-layer a horse standing in a field field a	**12-layer** a group of zebras drinking a a river	**12-layer** a group of people beach on a beach beach	**12-layer** a man on skis on on a a
15-layer a couple grazing in a a a a	**15-layer** a herd of zebras standing a a a	**15-layer** a people of people flying kites flying kites kites	**15-layer** a people of skis skis skis a a

and 9-layer GRU is selected as the optimal configuration. Then, the proposed 9-layer image captioning approach is compared with the state-of-the-art approaches as listed in Table II. It can be observed that our proposed approach outperforms the state-of-the-art approaches in terms of four metric scores. The approaches are sorted based on CIDEr metrics, and the highest score is indicated with bold fonts in each column.

Table III shows the reference captions and generated captions by the proposed approach for four images. From those results, we observe that our proposed approach is capable of capturing image information with correct and descriptive captions. For instance, in the first image, the generated caption can successfully describe the *horse* and *mountains* with their positions in the image. The proposed approach finds the *crossing* action in the second image whereas the *posing* action in the fourth image. In the third image, the proposed approach generates the words *beach* and *kite*, which accurately describe the content of the image. Results show that our proposed approach can generate natural sentences related to the image.

IV. CONCLUSION & FUTURE WORK

In this paper, a novel image captioning approach based on the Inception-v3 CNN encoder and multi-layer GRU based decoder has been presented. The empirical results on the MSCOCO dataset indicate that the proposed approach can generate semantically coherent captions by integrating a multi-layer GRU based decoder. Our proposed multi-layer GRU based decoder achieves competitive captioning performance compared with state-of-the-art methods on image captioning tasks. Our future work will focus on generating more accurate captions with higher CIDEr scores integrating the self-attention mechanism and transformer.

ACKNOWLEDGMENT

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 projects: 120N995, & 623805725) and by the scientific research projects coordination unit of Izmir Katip

Celebi University (project no: 2021-ÖDL-MÜMF-0006).

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*. IEEE, 2019, pp. 945–949.

[2] R. Keskin, Ö. Çaylı, Ö. T. Moral, V. Kılıç, and A. Onan, "A benchmark for feature-injection architectures in image captioning," *European Journal of Science and Technology*, no. 31, pp. 461–468, 2021.

[3] B. Fetiler, Ö. Çaylı, Ö. T. Moral, V. Kılıç, and A. Onan, "Video captioning based on multi-layer gated recurrent unit for smartphones," *European Journal of Science and Technology*, no. 32, pp. 221–226, 2021.

[4] S. Aydın, Ö. Çaylı, V. Kılıç, and A. Onan, "Sequence-to-sequence video captioning with residual connected gated recurrent units," *Avrupa Bilim ve Teknoloji Dergisi*, no. 35, pp. 380–386, 2022.

[5] B. Uslu, Ö. Çaylı, V. Kılıç, and A. Onan, "Resnet based deep gated recurrent unit for image captioning on smartphone," *Avrupa Bilim ve Teknoloji Dergisi*, no. 35, pp. 610–615, 2022.

[6] S. Li, G. Kulkarni, T. L. Berg, A. C. Berg, and Y. Choi, "Composing simple image descriptions using web-scale n-grams," in *Proceedings of the 15th Conference on Computational Natural Language Learning*. Association for Computational Linguistics, 2011, pp. 220–228.

[7] G. Kulkarni, V. Premraj, V. Ordonez, S. Dhar, S. Li, Y. Choi, A. C. Berg, and T. L. Berg, "Babytalk: Understanding and generating simple image descriptions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 12, pp. 2891–2903, 2013.

[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] 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*, 2014.

[10] B. Makav and V. Kılıç, "Smartphone-based image captioning for visually and hearing impaired," in *11th International Conference on Electrical and Electronics Engineering*. IEEE, 2019, pp. 950–953.

[11] 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*. Springer, 2020, pp. 1532–1539.

[12] Q. You, H. Jin, and J. Luo, "Image captioning at will: A versatile scheme for effectively injecting sentiments into image descriptions," *arXiv preprint arXiv:1801.10121*, 2018.

[13] 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.

[14] 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.

[15] A. Mathews, L. Xie, and X. He, "Semstyle: Learning to generate stylised image captions using unaligned text," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8591–8600.

[16] 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.

[17] Y. H. Tan and C. S. Chan, "Phrase-based image caption generator with hierarchical lstm network," *Neurocomputing*, vol. 333, pp. 86–100, 2019.

[18] X. Jia, E. Gavves, B. Fernando, and T. Tuytelaars, "Guiding the long-short term memory model for image caption generation," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 2407–2415.

[19] Q. Wang and A. B. Chan, "Cnn+ cnn: Convolutional decoders for image captioning," *arXiv preprint arXiv:1805.09019*, 2018.

[20] 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.

[21] J. Devlin, H. Cheng, H. Fang, S. Gupta, L. Deng, X. He, G. Zweig, and M. Mitchell, "Language models for image captioning: The quirks and what works," *arXiv preprint arXiv:1505.01809*, 2015.

[22] S. Liu, Z. Zhu, N. Ye, S. Guadarrama, and K. Murphy, "Optimization of image description metrics using policy gradient methods," *arXiv preprint arXiv:1612.00370*, vol. 5, 2016.

[23] O. Nina and A. Rodriguez, "Simplified lstm unit and search space probability exploration for image description," in *10th International Conference on Information, Communications and Signal Processing*. IEEE, 2015, pp. 1–5.

[24] 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.

[25] 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.

[26] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, "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.

[27] 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.

[28] 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 International Conference on Computer Vision*, 2015, pp. 2533–2541.

[29] M. Baran, Ö. T. Moral, and V. Kılıç, "Merge model based image captioning for smartphones," *European Journal of Science and Technology*, no. 26, pp. 191–196, 2021.

[30] R. Keskin, Ö. T. Moral, V. Kılıç, and A. Onan, "Multi-gru based automated image captioning for smartphones," in *2021 29th Signal Processing and Communications Applications Conference (SIU)*. IEEE, 2021, pp. 1–4.

[31] V. Kılıç, "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.

[32] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European Conference on Computer Vision*. Springer, 2014, pp. 740–755.

[33] 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.

[34] D. Gurari, Y. Zhao, M. Zhang, and N. Bhattacharya, "Captioning images taken by people who are blind," in *European Conference on Computer Vision*. Springer, 2020, pp. 417–434.

[35] 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*. Association for Computational Linguistics, 2002, pp. 311–318.

[36] 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.

[37] A. Lavie and A. Agarwal, "Meteor: An automatic metric for mt evaluation with high levels of correlation with human judgments," in *Proceedings of the Second Workshop on Statistical Machine Translation*, 2007, pp. 228–231.

[38] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Proceedings of the ACL-04 Workshop*. ACL, 2004, pp. 1–8.

[39] P. Anderson, B. Fernando, M. Johnson, and S. Gould, "Spice: Semantic propositional image caption evaluation," in *European Conference on Computer Vision*. Springer, 2016, pp. 382–398.

[40] 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 and Learning Systems*, vol. 32, no. 2, pp. 604–624, 2020.

[41] R. Staniūtė and D. Šešok, "A systematic literature review on image captioning," *Applied Sciences*, vol. 9, no. 10, p. 2024, 2019.