Dicle University
Journal of Engineering
https://dergipark.org.tr/tr/pub/dumf
duje.dicle.edu.tr

Research Article

From Pixels to Paragraphs: Exploring Enhanced Image-to-Text Generation using Inception v3 and Attention Mechanisms

Zeynep KARACA*, Bihter DAS[2]

[1] Firat University, Technology Faculty, Software Engineering Department, krczeynep1996@outlook.com, Orcid No: 0000-0002-7751-8567
[2] Firat University, Technology Faculty, Software Engineering Department, bihterdas@firat.edu.tr, Orcid No 0000-0002-2498-3297

ARTICLE INFO

Article history:

Received 10 August 2023
Received in revised form 4 November 2023
Accepted 11 November 2023
Available online 31 December 2023

Keywords:

Inception v3 Model, Attention Mechanisms, Textual Content Extraction, Image-to-Text Generation

Doi: 10.24012/dumf.1340656

* Corresponding author

ABSTRACT

Processing visual data and converting it into text plays a crucial role in fields like information retrieval and data analysis in the digital world. At this juncture, the "image-to-text" transformation, which bridges the gap between visual and textual data, has garnered significant interest from researchers and industry experts. This article presents a study on generating text from images. The study aims to measure the contribution of adding an attention mechanism to the encoder-decoder-based Inception v3 deep learning architecture for image-to-text generation. In the model, the Inception v3 model is trained on the Flickr8k dataset to extract image features. The encoder-decoder structure with an attention mechanism is employed for next-word prediction, and the model is trained on the train images of the Flickr8k dataset for performance evaluation. Experimental results demonstrate the model's satisfactory ability to accurately perceive objects in images.

Introduction

In today's digital age, the efforts towards converting visual content into text have continued to be a critical step for various applications such as data analysis, content indexing, search engines, and many more. This transformation allows visual data to become more accessible and meaningful by adopting a textual format [1]. At this juncture, deep learning techniques have played a pivotal role in the conversion of visual data into textual descriptions. Deep learning, a subset of artificial intelligence, revolves around neural networks and utilizes vast amounts of data to automatically extract intricate features [2]. These techniques have significantly enhanced the performance of the process of transforming images into text. The integration of deep learning into the image-to-text conversion process has resulted in substantial performance improvements. Particularly, encoder-decoder architectures combine convolutional neural networks (CNNs) to extract visual features and recurrent neural networks (RNNs) to translate these features into textual descriptions. This fusion enables the effective capture of essential features and patterns within images, ultimately leading to the generation of textual explanations that are natural and coherent. Moreover, the use of techniques such as attention mechanisms contributes to further refining image-to-text conversion [3]. Attention mechanisms emphasize specific visual features during text generation, leading to more meaningful and contextually rich descriptions. By enhancing image-to-text transformation performance, attention mechanisms contribute to achieving consistent and high-quality outcomes. Supported by deep learning techniques, the conversion of visual data to textual descriptions stands as a pivotal step, facilitating the generation of more meaningful, consistent, and natural texts. Through these approaches, a broader spectrum of applications that utilize visual data in text format can be realized [4].

The rest of the paper has been organized as follows: In the 2nd section, related studies from the literature are presented. The 3rd section details the dataset specific to the proposed model, the InceptionV3 method, and the Attention Mechanism. In the 4th section, Experimental Results and Discussion are provided. Finally, the last section contains the Conclusion of the study.

Literature Review

In the literature, various studies have been conducted on generating text from images, most of which have utilized machine learning and deep learning techniques.

Kanimozhiselvi et al. utilized three CNN architectures, namely Inception-V3, ResNet50, and Xception models, for feature extraction, and employed LSTM for caption generation. They used the Flickr 8k dataset and achieved the highest accuracy with the Xception architecture, obtaining a 75% accuracy after training Xception + LSTM for 50 epochs [5]. Bai et al. employed a CNN-based generation model using Conditional Generative Adversarial Networks (CGAN) to create image captions. They used Multi-modal Graph Convolutional Networks (MGCN) to establish visual relationships between objects. Their experiments on the MSCOCO 2014 dataset showed improved performance compared to state-of-the-art methods [6]. Agrawal et al. proposed a model based on encoder and attention-based decoder. They used a pre-trained Convolutional Neural Network (CNN) as the encoder and introduced an attention mechanism to generate captions that best match the image. They employed the Inception v3 architecture and Recurrent Neural Networks (RNN) to extract image features and create captions. The model incorporated the Bahdanau Attention Mechanism and performed better compared to traditional methods [7]. Kılıçkaya et al. addressed the problem using the Im2Text method focusing on meta-class features. They used the Pascal Sentences dataset consisting of 1000 images, each associated with 5 different captions created by 5 individuals, totaling 5000 captions. Their approach yielded a Bleu1 score of 0.0067 [8]. Lu et al. aimed to generate captions for fine art images by developing a virtual-real semantic alignment training process. They used the MS COCO and ArtCap datasets during model training. Their model achieved a Bleu1 performance of 0.508, a Meteor performance of 0.1317 [9]. Yang et al. focused on generating human-centric captions to determine human behaviors. They introduced the Human-Centric Caption Model (HCCM) relying on detailed feature extraction and interaction. They proposed a three-branch hierarchical caption model, creating a dataset called Human-Centric COCO (HC COCO). While showing improvements over existing methods, their approach fell short in generating detailed captions [10]. Li et al. proposed a semantic matching method combining semantic similarities to learn hidden correlations between images and captions. They used local semantic similarity measurement mechanisms based on the comparison of semantic units. The model achieved high-performance results on the MSCOCO dataset, with Bleu1 at 81.2, Bleu4 at 39.0, Rouge_l at 58.9, and CIDEr-D at 128.5 [11]. Jaknamon et al. presented a Transformer-based approach named ThaiTC for image captioning. They used image transformer and text transformer instead of traditional CNN and RNN for encoding and decoding, respectively. Their experiments showed varying performances on different datasets [12]. Krisna et al. aimed to generate effective and accurate captions for rainy-noisy images. They developed an end-to-end architecture using GAN-based methods. They incorporated a conditional GAN architecture for handling distorted images, an Inception v3 encoder, and a Bahdanau Attention mechanism-based GRU decoder. Their model performed well in generating captions [13]. Shambharkar et al. proposed a beam-search based CNN+RNN architecture which generates multiple captions for an image and selects the best caption among these based on similarity with reference captions. The RSCID dataset was utilized, and their approach outperformed the non-beam-searched encoder-decoder architecture [14]. Feng et al. proposed a model that integrates caption and gaze tracking by learning the relationship between captions and eye-tracking patterns. The dataset contained 400 training images, 200 validation images, and 400 test images. The model showed a Recall@5 performance of 0.0048 [15]. Cai et al. presented a multimodal fashion image captioning model, achieving a performance of Bleu1 at 46.5, Meteor at 22.3, and Rouge at 38.6 [16]. Ye et al. proposed a joint training two-stage (JTTS) method for remote sensing image captions. They used RSICD, UCM-captions, and Sydney-captions datasets, achieving high-performance results [17]. Wang et al. introduced a caption transformer (CapFormer) architecture for remote sensing image captions. Their model achieved performance improvements, with Bleu1 at 66.12 and Rouge_l at 49.78 [18]. Malhotra et al. proposed a model using ResNet50 for image encoding and RNN and LSTM for sentence generation, achieving F1 Score 77.8, Meteor 27.6, and accuracy 70 [19]. Yang et al. proposed the Context-Sensitive Transformer Network (CSTNet) method, achieving improved performance compared to SOTA, with Bleu1 at 81.1, Meteor at 29.4, and Rouge at 59.0 [20]. Wang et al. suggested a parallel fusion RNN+LSTM architecture that enhances efficiency and achieves better results than the dominant approach. Their model obtained Bleu1 at 66.7 and Meteor at 16.53 after training [21]. Table 1 shows a comparison of the studies.

Table 1. The comparison of literature reviews

Author	**Method**	**Dataset**	**Bleu-1**	**Bleu-4**	**Meteor**	**Rouge-L**	**Cider**
Kılıçkaya et al.[8]	Im2Text	Pascal Sentences	0.0067				
Lu et al.[9]	Semantic Alignment method	MSCOCO ArtCap	0.508		0.1317		
Li et al. [11]	Semantic Matching	MSCOCO	81.2	39.0		58.9	128.5
Shambharkar et al. [14]	Beam-Search CNN+RNN	RSCID					

Cai et al. [16]	Multimodal fashion	FACAD	46.5		22.3	38.6
Ye et al. [17]	JTTS	UCM-Captions	0.8696			0.8364
		Sydney-Captions	0.8492			0.7660
		RSICD	0.7893			0.6823
Wang et al. [18]	CapFormer	RSICD+ GoogleEarth	66.12			48.78
Yang et al. [20]	CSTNet	MSCOCO	81.1		29.4	59.0
Wang et al. [21]	RNN+LSTM	Flikr8k + Neural Talk1	66.7		16.53	
The proposed	InceptionV3	Flickr8k	52.77	48.39	64.17	89.90

Material and Methods

In this study, a deep learning architecture for automatic image caption generation is presented by incorporating an attention mechanism. In our application, the Inception v3 deep learning architecture with an encoder-decoder architecture is utilized. The Inception v3 architecture serves as the encoder, with a CNN-based encoder and an LSTM-based decoder. To enhance the meaningfulness and detail of generated captions, an attention mechanism is integrated into the Inception v3 model. Through the CNN-based encoder, a thought vector is obtained, extracting image features. The generated image features and image details are then fed into the LSTM decoder. The LSTM decoder produces captions in natural language. The added attention mechanism in the model oversees the input sequence during each step of the decoding process, focusing on selected portions of the input sequence at specific steps. This mechanism contributes to creating captions that better describe the image. The Inception v3+Attention Mechanism model is trained using the Flickr8k dataset. The proposed architecture is illustrated in Figure 1.

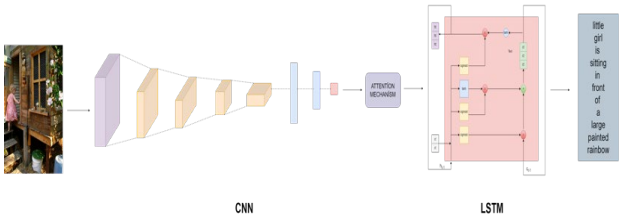

Figure 1. Architecture of the proposed system

Dataset

In this study, the Flickr8k dataset was utilized. The Flickr8k dataset comprises multiple captions for each image. The dataset contains a total of 8091 images. The textual files within the dataset provide descriptions of objects and events present in the images. Each image is associated with five captions. The inclusion of multiple captions for a single image aids in generalizing the model and facilitating more accurate caption generation [24]. Figure 2 showcases some images from the Flickr8k dataset, while Figure 3 displays an image along with its corresponding caption from the Flickr8k dataset.

Figure 2. Some of images in the Flickr8k

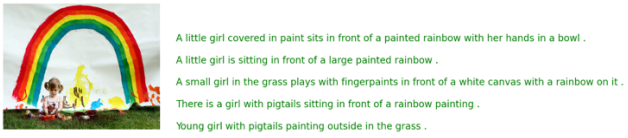

Figure 3. An example for image-to-text from the Flickr8k

Convolutional Neural Network

Convolutional Neural Networks (CNNs) are a class of deep learning architectures that have been successfully employed in various fields such as image recognition, image localization, speech recognition, language translation, and image Classification [22]. The CNN architecture is commonly utilized in the domain of image processing, where it takes an image as input and transforms it into a

matrix format for recognition and processing by computers. The CNN method consists of different layers that extract and classify features from diverse aspects of images.

Starting with the input image, it undergoes Convolutional Layer, Pooling Layer, and Fully Connected Layer operations, ultimately being prepared as input for the deep learning model. The Convolutional Layer is the first layer that processes the image in CNN architectures. It employs a filter smaller than the image's dimensions to extract features. The model updates filter values iteratively to enhance the extraction of image features. The Pooling Layer focuses on discarding irrelevant features and concentrating on more important ones. Two common techniques used in CNN architectures are Average Pooling and Max Pooling. The filter employed in the convolutional layer is also used in the pooling layer, depending on the chosen pooling method [22]. For max pooling, the maximum value in the filter region is selected, while for average pooling, the average of the values in the filter region is taken. After passing through the convolutional and pooling layers, the image is transformed into a matrix and then further into a vector through the Fully Connected Layer. The input image that traverses through these layers becomes ready for training with neural networks [23]. Figure 4 shows the used CNN architecture.

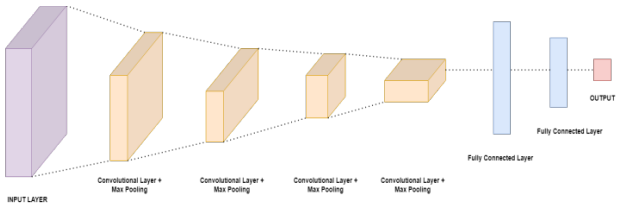

Figure 4. The CNN architecture

Inception V3 Model

The Inception V3 deep learning architecture, developed by Google, is a multi-level feature extraction image recognition model. The Inception V3 deep learning architecture employs Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) units. The Inception V3 model with a CNN-based architecture has been trained on over a million images from the ImageNet dataset [24]. It is referred to as Inception v3 due to the existence of different versions of this architecture. The key difference from previous architectures lies in the addition of convolutional layers, batch normalization, and fully connected layers [25]. It achieves a lower error rate compared to previous versions and similar models.

The Inception v3 deep learning architecture extracts useful features from input images in the training part of the dataset and utilizes these features more effectively [26]. Unlike Inception v1 and Inception v2, Inception v3 has more layers, consisting of 42 layers. Despite its 42-layer depth, it offers better accuracy and lower computational costs compared to the previous version. Despite having 42 layers, it is 2.5 times more efficient than GoogleNet and more efficient than AlexNet [27]. The Inception v3 architecture is an image recognition model that achieves more than 78.1% accuracy on the ImageNet dataset [28]. The Inception v3 model comprises 21 million parameters. The input images for the Inception v3 model are in the shape of 299 x 299 x 3 (RGB). The layer structure of the model is illustrated in Figure 5.

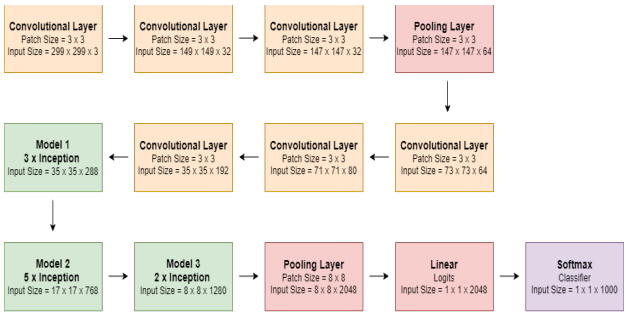

Figure 5. The used Inception V3 architecture

Long-Short Term Memory

LSTM was designed to overcome the short-term memory issues of RNNs by providing better memory storage. Due to its ability to capture long-term dependencies and preserve sequences, LSTM has become the most widely used RNN architecture [29]. LSTM finds applications in various fields such as language modeling, text generation, image processing, handwriting recognition, music processing, and text translation. It has the capability to retain information over long periods [30]. LSTM utilizes a cell state and various gates. The cell state is a pathway that carries information and enables predictions, often referred to as the memory of the network. The decision about what information to carry along the cell state is made by gates. These gates determine whether certain pieces of information are necessary or not. Figure 6 shows the LSTM structure.

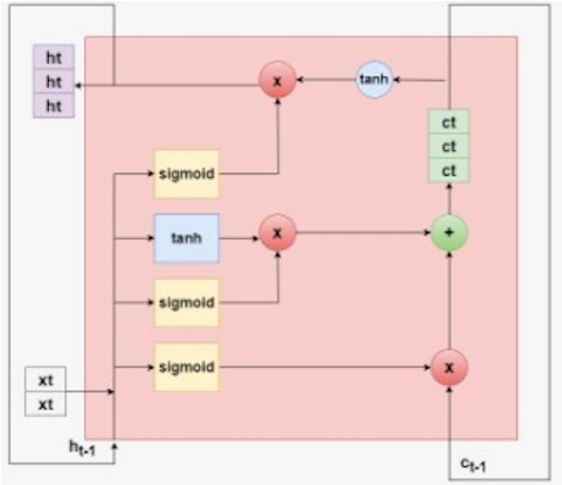

Figure 6. The LSTM structure

Attention Mechanism

The attention mechanism aims to create more focused captions that encompass the highlights of an image [7]. Visual attention is a mechanism that filters out irrelevant information from complex visual scenes. When generating a caption, the model ensures that it pays more attention to relevant areas in the image [28].

In sequence-to-sequence models, the entire input is compressed into a fixed-length vector through the encoder to be fed into the decoder. The intuition behind the Attention Mechanism is that rather than compressing the input, it might be better for the decoder to look back at the input sequence at each step. The decoder can take the context vector, which is a weighted sum of representations of the input data at each step, as input. Intuitively, the weights determine how much focus each step's content should place on the input token, and these weights can be differentiated to learn the attention mechanism alongside other neural network parameters. Models that incorporate attention mechanisms demonstrate better performance compared to the original sequence-to-sequence architectures [28].

Encoder-decoder

The input image is introduced to the system using the Inception V3 model. The input image is sent to the CNN (Convolutional Neural Network) part of the Inception V3 model. The features of the image are extracted by the encoder. CNN processes the images and generates an image feature map. The attention mechanism takes these extracted feature maps along with the hidden state. A weight is assigned to each pixel of the image. These weights are then combined with the input words at each time step. Once combined, they are sent to the LSTM (Long Short-Term Memory) network. The decoder takes the features extracted by CNN as input and generates captions using LSTM. The image features, which are a fixed-length vector, feed into the decoder as input along with the token indicating the beginning of the sequence, <start>. To produce the output word of the decoder, at each time step, it takes the previous time step's hidden state and the predicted word. This process is repeated for all words until the sequence end token, <end>, is generated.

The decoder incorporates an attention mechanism to generate the created captions in a more detailed and accurate manner. Thanks to the attention mechanism, the captions are produced in a more focused manner. The attention mechanism focuses on the region of the image and the previously generated words to create captions.

Experimental Results

Various automatic evaluation criteria are used to assess the success of text generation in the literature. Popular metrics for evaluating image-to-text generation include BLEU, ROUGE, METEOR, and CIDer. These metrics produce values between 0-1 or 0-100. Lower values indicate that the generated captions poorly describe the image, while higher values suggest that the captions effectively describe the image and are as good as human translations.

- Bi-Lingual Evaluation Understudy (BLEU): Compares machine-generated translation to human-generated translation regardless of word order. Calculates by comparing word matches and dividing candidate words by reference words. Ranges from 0 to 1, with higher values indicating better alignment.
- Recall-Oriented Understudy for Gisting Evaluating (ROUGE): An N-gram based metric comparing N-grams between machine and reference translations. ROUGE metrics provide scores between 0 and 1. Higher scores near 1 suggest strong matching between generated and reference captions.
- ROUGE_L: It measures longest matching sequence of words.
- ROUGE_N: It measures unigram, bigram, trigram and higher order n-gram overlap.
- Metric for Evaluation of Translation with Explicit Ordering (METEOR): METEOR adjusts precision and recalls calculations.

Table 2 shows the performance results of the proposed system using Inception V3 and the Attention mechanism.

Table 2. The performance results.

Metrics	Results
BLEU-1	52,77
BLEU-2	50,63
BLEU-3	49,76
BLEU-4	48,39
METEOR	64,17
ROUGE_L	**89,90**
ROUGE_N	**81,88**

According to Table 1, it is seen that the text similarity and n-gram text similarity using the longest common subsequences are high. Figure 7 shows the results of the real captions and prediction captions.

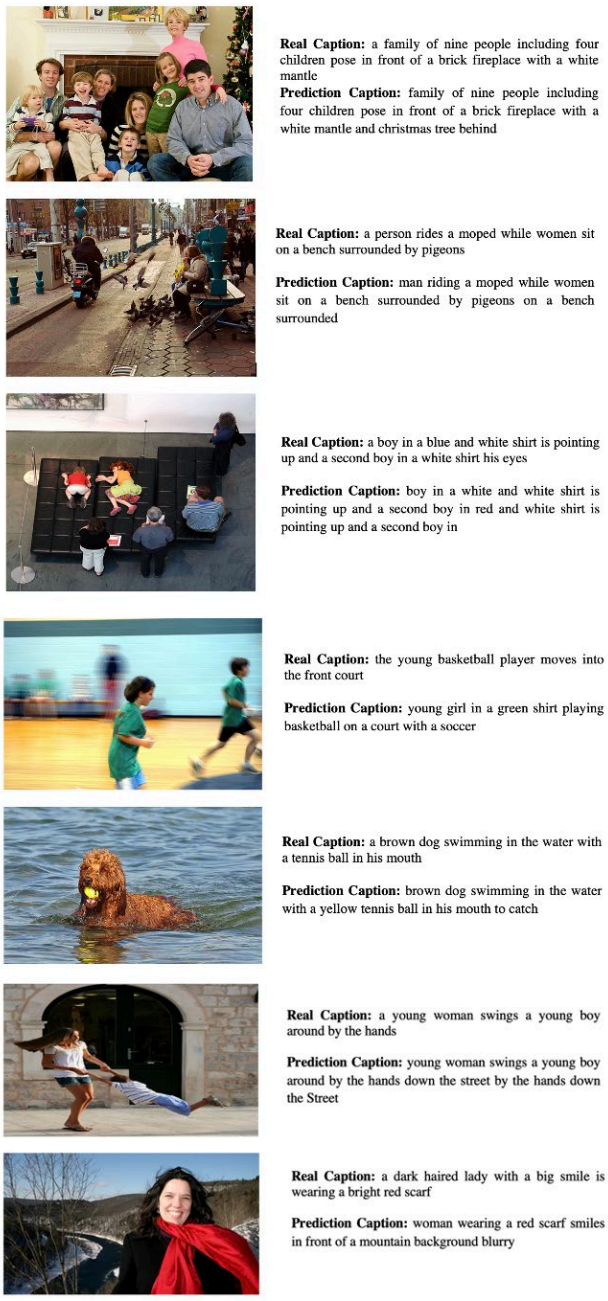

Figure 7. Results of prediction captions

According to Figure 7, it is seen that the captions produced with the proposed system better express the real pictures and draw attention not only to the people or colors in the image but also to the details in the background. It is clear that the proposed system pays better attention to the details in the image. For example, in the image of the girl wearing a red scarf, the real caption does not pay attention to the details in the background, whereas the prediction made with the proposed system states that the girl is both wearing a red scarf and is in front of a mountain in the back, and the image is blurry.

Conclusion

In this study, existing approaches in the literature for obtaining image captions were reviewed, and an approach that incorporates the attention mechanism in addition to the Inception V3 model was proposed. In the proposed approach, images extracted from the Flickr8k dataset were provided as input to the model, aiming to generate a sentence that describes the given image as the output. The images were fed into the Inception V3 model, and a CNN+LSTM based model with an added attention mechanism was employed. The images were processed through the CNN to extract image feature maps. The attention mechanism took these features as hidden states. The decoder then utilized the image features to generate captions. The generated captions for each image have accurately perceived the objects in the images, resulting in coherent and satisfying sentences.

Results of the ROUGE_L, ROUGE_N, and METEOR metrics were close to 1, indicating high performance. While the BLEU-1, BLEU-2, BLEU-3, and BLEU-4 scores are not very low, they appear to be relatively lower compared to the other metrics, which could be attributed to the small dataset used in this study. This proved to be a disadvantage for our work. In future research, improvements in background modeling and the utilization of larger datasets are planned to generate more meaningful captions for images.

Acknowledgment

This study was supported by Fırat University Scientific Research Projects Coordination Unit (FUBAP) with the project number ADEP.22.06."

References

[1] M. Bahani, A. E. Ouaazizi, and K. Maalmi, "The effectiveness of T5, GPT-2, and BERT on text-to-image generation task," Pattern Recognition Letters, Aug. 2023, doi: 10.1016/j.patrec.2023.08.001.

[2] Y. Tian, A. Ding, D. Wang, X. Luo, B. Wan, and Y. Wang, "Bi-Attention enhanced representation learning for image-text matching," Pattern Recognition, vol. 140, p. 109548, Aug. 2023, doi: 10.1016/j.patcog.2023.109548.

[3] H. Polat, M. U. Aluçlu, and M. S. Özerdem, "Evaluation of potential auras in generalized epilepsy from EEG signals using deep convolutional neural networks and time-frequency representation," Biomedical Engineering / Biomedizinische Technik, vol. 65, no. 4, pp. 379-391, 2020, doi: 10.1515/bmt-2019-0098.

[4] H. Elfaik and E. H. Nfaoui, "Leveraging feature-level fusion representations and attentional bidirectional RNN-CNN deep models for Arabic affect analysis on Twitter," Journal of King Saud University - Computer and Information Sciences, vol. 35, no. 1, pp. 462–482, Jan. 2023, doi: 10.1016/j.jksuci.2022.12.015.

[5] C. S. Kanimozhiselvi, K. V, K. S. P, and K. S, "Image Captioning Using Deep Learning," in 2022 International Conference on Computer Communication and Informatics (ICCCI), Jan. 2022, pp. 1-7, doi: 10.1109/ICCCI54379.2022.9740788.

[6] C. Bai, A. Zheng, Y. Huang, X. Pan, and N. Chen, "Boosting convolutional image captioning with

semantic content and visual relationship," Displays, vol. 70, p. 102069, Dec. 2021, doi: 10.1016/j.displa.2021.102069.

[7] V. Agrawal, S. Dhekane, N. Tuniya, and V. Vyas, "Image Caption Generator Using Attention Mechanism," in 2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT), Jul. 2021, pp. 1-6, doi: 10.1109/ICCCNT51525.2021.9579967.

[8] M. Kılıçkaya, E. Erdem, A. Erdem, N. İ. Cinbiş, and R. Çakıcı, "Data-driven image captioning with meta-class based retrieval," in 2014 22nd Signal Processing and Communications Applications Conference (SIU), Apr. 2014, pp. 1922-1925, doi: 10.1109/SIU.2014.6830631.

[9] Y. Lu, C. Guo, X. Dai, and F.-Y. Wang, "Data-efficient image captioning of fine art paintings via virtual-real semantic alignment training," Neurocomputing, vol. 490, pp. 163-180, Jun. 2022, doi: 10.1016/j.neucom.2022.01.068.

[10] Z. Yang, P. Wang, T. Chu, and J. Yang, "Human-Centric Image Captioning," Pattern Recognition, vol. 126, p. 108545, Jun. 2022, doi: 10.1016/j.patcog.2022.108545.

[11] J. Li, N. Xu, W. Nie, and S. Zhang, "Image Captioning with multi-level similarity-guided semantic matching," Visual Informatics, vol. 5, no. 4, pp. 41-48, Dec. 2021, doi: 10.1016/j.visinf.2021.11.003.

[12] T. Jaknamon and S. Marukatat, "ThaiTC:Thai Transformer-based Image Captioning," in 2022 17th International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP), Nov. 2022, pp. 1-4, doi: 10.1109/iSAI-NLP56921.2022.9960246.

[13] A. Krisna, A. S. Parihar, A. Das, and A. Aryan, "End-to-End Model for Heavy Rain Image Captioning," in 2022 4th International Conference on Advances in Computing, Communication Control and Networking (ICAC3N), Dec. 2022, pp. 1646-1651, doi: 10.1109/ICAC3N56670.2022.10074181.

[14] P. G. Shambharkar, P. Kumari, P. Yadav, and R. Kumar, "Generating Caption for Image using Beam Search and Analyzation with Unsupervised Image Captioning Algorithm," in 2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS), May 2021, pp. 857-864, doi: 10.1109/ICICCS51141.2021.9432245.

[15] Y. Feng, K. Maeda, T. Ogawa, and M. Haseyama, "Human-Centric Image Retrieval with Gaze-Based Image Captioning," in 2022 IEEE International Conference on Image Processing (ICIP), Oct. 2022, pp. 3828-3832, doi: 10.1109/ICIP46576.2022.9897949.

[16] C. Cai, K.-H. Yap, and S. Wang, "Attribute Conditioned Fashion Image Captioning," in 2022 IEEE International Conference on Image Processing (ICIP), Oct. 2022, pp. 1921-1925, doi: 10.1109/ICIP46576.2022.9897417.

[17] X. Ye et al., "A Joint-Training Two-Stage Method For Remote Sensing Image Captioning," IEEE Transactions on Geoscience and Remote Sensing, vol. 60, pp. 1-16, 2022, doi: 10.1109/TGRS.2022.3224244.

[18] J. Wang, Z. Chen, A. Ma, and Y. Zhong, "Capformer: Pure Transformer for Remote Sensing Image Caption," in IGARSS 2022 - 2022 IEEE International Geoscience and Remote Sensing Symposium, Jul. 2022, pp. 7996-7999, doi: 10.1109/IGARSS46834.2022.9883199.

[19] R. Malhotra, T. Raj, and V. Gupta, "Image Captioning and Identification of Dangerous Situations using Transfer Learning," in 2022 6th International Conference on Computing Methodologies and Communication (ICCMC), Mar. 2022, pp. 909-915, doi: 10.1109/ICCMC53470.2022.9753788.

[20] Xin Yang et al., "Context-Aware Transformer for image captioning," Neurocomputing, vol. 549, p. 126440, 2023, doi: 10.1016/j.neucom.2023.126440.

[21] M. Wang, L. Song, X. Yang, and C. Luo, "A parallel-fusion RNN-LSTM architecture for image caption generation," in 2016 IEEE International Conference on Image Processing (ICIP), Sep. 2016, pp. 4448-4452, doi: 10.1109/ICIP.2016.7533201.

[22] M. Şeker and M. S. Özerdem, "Automated Detection of Alzheimer's Disease using raw EEG time series via. DWT-CNN model," Dicle Üniversitesi Mühendislik Fakültesi Mühendislik Dergisi, vol. 13, no. 4, pp. 673-684, Jan. 2023, doi:10.24012/dumf.1197722.

[23] S. Örenç, E. Acar, and M. S. Özerdem, "Utilizing the Ensemble of Deep Learning Approaches to Identify Monkeypox Disease," Dicle Üniversitesi Mühendislik Fakültesi Mühendislik Dergisi, vol. 13, no. 4, pp. 685-691, Jan. 2023, doi:10.24012/dumf.1199679.

[24] S. Degadwala, D. Vyas, H. Biswas, U. Chakraborty, and S. Saha, "Image Captioning Using Inception V3 Transfer Learning Model," in 2021 6th International Conference on Communication and Electronics Systems (ICCES), Jul. 2021, pp. 1103-1108, doi: 10.1109/ICCES51350.2021.9489111.

[25] O. Turk, D. Ozhan, E. Acar, T. C. Akinci, and M. Yilmaz, "Automatic detection of brain tumors with the aid of ensemble deep learning architectures and class activation map indicators by employing magnetic resonance images," Zeitschrift für Medizinische Physik, Dec. 2022, doi: 10.1016/j.zemedi.2022.11.010.

[26] K. Joshi, V. Tripathi, C. Bose, and C. Bhardwaj, "Robust Sports Image Classification Using InceptionV3 and Neural Networks," Procedia

Computer Science, vol. 167, pp. 2374-2381, Jan. 2020, doi: 10.1016/j.procs.2020.03.290.

[27] C. Szegedy et al., "Rethinking the Inception Architecture for Computer Vision," in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Jun. 2016, pp. 2818-2826, doi: 10.1109/CVPR.2016.308.

[28] X. Yu, Y. Ahn, and J. Jeong, "High-level Image Classification by Synergizing Image Captioning with BERT," in 2021 International Conference on Information and Communication Technology Convergence (ICTC), Oct. 2021, pp. 1686-1690, doi: 10.1109/ICTC52510.2021.9620954.

[29] C. Zhang, Y. Dai, Y. Cheng, Z. Jia, and K. Hirota, "Recurrent Attention LSTM Model for Image Chinese Caption Generation," in 2018 Joint 10th International Conference on Soft Computing and Intelligent Systems (SCIS) and 19th International Symposium on Advanced Intelligent Systems (ISIS), Dec. 2018, pp. 808-813, doi: 10.1109/SCIS-ISIS.2018.00134.

[30] K. Xu, H. Wang, and P. Tang, "Image captioning with deep LSTM based on sequential residual," in 2017 IEEE International Conference on Multimedia and Expo (ICME), Jul. 2017, pp. 361-366, doi: 10.1109/ICME.2017.8019408.