Proceedings of the Eighth International Conference on Research in Intelligent Computing in Engineering pp. 105–110
DOI: 10.15439/2023R55
ISSN 2300-5963 ACSIS, Vol. 38

DICKT - Deep Learning-Based Image Captioning using Keras and TensorFlow

Phung Thao Vi
International School – Vietnam
National University
Hanoi, Vietnam
phungvi08123@gmail.com

Satyam Mishra
International School – Vietnam
National University
Hanoi, Vietnam
satyam.entrprnr@gmail.com
0000-0002-7457-0060

Le Anh Ngoc*
Swinburne Vietnam,
FPT University
Hanoi, Vietnam
ngocla2@fpt.edu.vn

Sundaram Mishra
NETMONASTERY NSPL, Mumbai, India
mishrasundaram.sm@gmail.com

Vu Minh Phuc
International School – Vietnam National University
Hanoi, Vietnam
vphuc2411@gmail.com

***Abstract*—This study evaluates a caption generation model's performance using the BLEU Score metric. The model generates descriptions for images, compared to reference captions with single and dual references. Results show a high BLEU Score, suggesting human-like captions. However, BLEU primarily measures linguistic similarity and n-gram overlap, missing full human-generated caption richness. The findings reveal the model's potential to convey image essence in text, but highlight BLEU Score limitations. TensorFlow and Keras are used for model development, acknowledging their widespread use but also their limitations. The research offers insights into caption generation model capabilities and urges a broader perspective on caption quality beyond quantitative metrics. While higher BLEU Scores are generally preferred, a "good" score varies with dataset and context. The study emphasizes a need for a more comprehensive approach to assess the quality and creativity of machine-generated captions.**

***Index Terms*—Image Captioning, Deep Learning, Keras, TensorFlow, BLEU.**

I. Introduction

The digital age, widespread amounts of photo information are generated day by day, from social media systems to surveillance structures. These pictures hold worthwhile facts, however having access to their content material stays a project. Image captioning, [1] the process of robotically producing descriptive textual descriptions for pics, bridges this gap and finds programs in content material retrieval, accessibility, and assistive technologies. Deep gaining knowledge has revolutionized photo captioning, permitting the development of greater accurate and contextually conscious structures [2]. In this research, authors delve into the realm of deep learning-primarily based picture captioning, leveraging Keras and TensorFlow, with a focus on technique, experiments, and implications.

Image captioning is a multi-modal [3] undertaking that mixes pc vision and herbal language processing (NLP). Authors look into using convolutional neural networks (CNNs) for picture feature extraction and recurrent neural networks (RNNs) with interesting mechanisms for producing coherent and contextually applicable captions. Through an extensive exploration of information series, preprocessing, model structure, education strategies, and evaluation metrics, authors aim to push the bounds of image captioning abilities. [4], [5]

This study contributes to the continuing improvement of smart structures capable of information and describing visual content, with capability packages in image search engines, automated content material tagging, and accessibility for the visually impaired. By addressing the technical demanding situations and barriers related to deep studying-primarily based photo captioning, authors offer treasured insights into the modern-day country of the artwork and open avenues for destiny studies in this rapidly evolving discipline. [6]

II. Literature Review

A. Image captioning and its applications

Image captioning is a field that combines herbal language processing (NLP) and laptop vision (CV) to generate textual descriptions for pictures. It makes use of strategies such as Long Short-Term Memory (LSTM) and Convolutional Neural Networks (CNN) to understand and interpret images, allowing the machine to offer captions in a commonplace language [7]. This generation has various applications, along with supporting blind individuals by way of presenting braille legible captions for photos, aiding in medical document generation, and automating the workflow of healthcare experts [8], [9]. However, the heavy computational burden and massive memory storage required by way of present captioning fashions restrict their deployment on aid-restricted gadgets. To address this, lightweight image captioning models were proposed, leveraging techniques which include compact function extraction and optimized go-modal fusion, resulting in decreased model size and advanced inference speed [10] . These advancements make image captioning models extra sensible for actual-international programs.

*Corresponding author

©PTI 2023

B. *Deep learning for Image Captioning*

Deep studying techniques are being used for photo captioning, which aims to generate descriptive and correct textual descriptions for pics. This technique involves leveraging convolutional neural networks (CNNs) for image function extraction and recurrent neural networks (RNNs) for sequential language generation. The procedure of picture captioning normally involves an photo encoder, inclusive of a CNN, to extract high-stage capabilities, and a language decoder, including an RNN, to generate captions [11], [12]. Deep studying has proven splendid success in diverse computer imaginative and prescient tasks, consisting of photograph captioning [13]. The use of deep neural networks enables computer systems to recognize and interpret visible content material, bridging the gap between the visual and textual domain names [14]. Image captioning has applications in diverse fields, which includes self-riding automobiles, robotics, and image analysis [15]. The improvement of accurate image captioning structures can make contributions to advancements in photograph expertise and human-gadget interplay with visual data.

C. *Previous approaches and State-of-the-Art*

Image captioning research has seen fast progress, from template-based total fashions to deep neural community models the use of the encoder-decoder structure. One success method is the usage of feature vectors extracted from place proposals received from an object detector. The Object Relation Transformer improves image captioning with the aid of incorporating data approximately the spatial relationship between detected objects via geometric interest [16]. In the field of biomedical picture captioning, there may be a want for assisting clinicians in the prognosis technique. Surveys have been conducted on biomedical photograph captioning, discussing datasets, assessment measures, and contemporary methods [17] [18]. A simple cosine similarity measure using the Mean of Word Embeddings (MOWE) of captions has shown high overall performance in unsupervised caption evaluation. The proposed metric WEbSim outperforms complex measures and units a brand-new baseline for caption assessment [19].

III. Methodology

First step is to import these libraries: Numpy, TensorFlow, Matplotlib.pyplot, and Pandas for our project.

After that authors will import specific modules:

keras.applications.vgg16: Importing the VGG16 model, which is a popular deep learning model for image classification.

- keras.applications.resnet50: Importing the ResNet50 model, another popular deep learning model for image classification.
- preprocess_input and decode_predictions: Functions for preprocessing images and decoding model predictions.
- keras.preprocessing: Modules for preprocessing data for deep learning models.
- keras.models: Modules for defining and working with neural network models.
- keras.layers: Modules for defining layers in neural networks.
- keras.layers.merge: Importing the add function for merging layers.

A. *Data Collection*

The data authors have used in our study is “Flickr8l” dataset - the "Flickr8k" dataset is a popular and widely used dataset in the field of computer vision and natural language processing (NLP). It is specifically used for tasks related to image captioning. Here's an overview of the dataset, it involves 6000 trainImage, 1000 devImages, and 1000 testImages.

Each image contains 5 captions that will be presenting that image. Here in figure 1 is an example of images and its unique identify names:

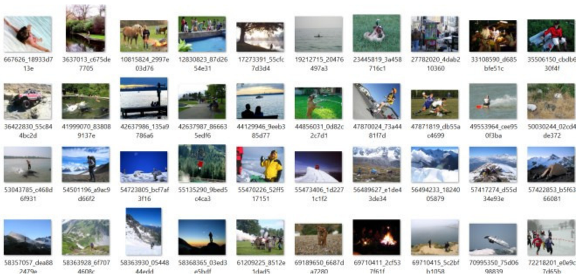


Fig.1. Image Caption Bot dataset

Each picture consists of descriptions of image content, in which every line represents a distinct photograph with a completely unique identifier and multiple descriptions. The descriptions are associated with the content material of the snapshots, offering facts about approximately what's happening or what is depicted in each photograph. is presenting a descriptions dictionary where key is ‘img_name’ and value is ‘cap’:

After reading the file and preprocessing the text file that contained images; it got the length of text, which is 40,460 lines (captions). A description dictionary that contains the image's captions is created, first is a list and after that obtaining captions one by one.

B. *Data Cleaning*

In the data cleaning process, the code follows a three-step approach:

1. Convert every word to its lowercase form.
2. Eliminate punctuations, unnecessary numbers, and special characters from the text.
3. Remove words with a length less than or equal to one.

This process results in clean sentences that effectively represent the associated images. To ensure efficiency and avoid redundancy when restarting the notebook, these cleaned sentences are stored in a dictionary, allowing us to access them directly at any time.

Next step is to access the world's frequency and features; it sorts words with a frequency less than a specified value. This step aims to eliminate outliers and uncommon words, retaining only those words that are frequently used to describe the data or depict images. Specifically, words must keep with a frequency greater than or equal to a threshold value of 10, as these words play a significant role in the model's purpose.

C. Loading of Training Set

For the loading of the training set, firstly, importing of the school dataset into our chosen development environment (e.g., Python with TensorFlow and Keras) for further processing. This step involves cleaning each caption and creating a dictionary. The '.jpg' characters are removed and keep only the image's names. These names with their corresponding captions are associated, it creates a 'trained_description' dictionary.

D. Data preprocessing

1) Data Preprocessing (Image)

In this section, images are loaded and do some processing so that it can feed it in the network. After this all, authors create two functions: one for loading an image and the other for converting it into an axis to enhance image features. Then the bottleneck features of the training store to disk. Then authors implement an encoding dictionary and an 'encode_image' function.

2) Data Preprocessing (For Image Caption)

The outcome of time is approximately 1818.58 seconds, it is the total time it took for encoding, and it is the important metric for evaluating performance and speed. Additionally, it will also be useful for tracking the performance and the efficiency of software and systems that have been used to process these images.

Next step, authors set up two dictionaries: 'word_to_index' and 'idx_to_word' . Since the input words cannot go directly into the model, the solution is that, authors convert words into corresponding indices. After prediction, numerical representations are obtained, which determine the maximum length of captions. For caption preprocessing, add 'Startseq' and 'Endseq' markers, and caption in the middle. The dataset employed in this study is extensive, as can be seen above, the predictive process initiates by introducing a "startseq" token as the first word. Subsequently, the model predicts the second word, which is appended to the "startseq" token. The process is repeated iteratively, with each predicted word being added to the sequence, and the model's predictions are influenced by both the image features and the preceding words. In essence, this progressive approach generates captions word by word.

3) Data Preparation using Generator Function

This data generator function is designed to provide input and output statistics in batches for schooling a neural community version for image captioning. It methods photograph descriptions, tokenizes them, and prepares them as enter-output pairs for the model. The generator keeps yielding batches of facts indefinitely until stopped externally.

E. Word Embeddings

To facilitate the manipulation of textual data, word embeddings are employed, transforming individual words into 50-dimensional vectors. These embeddings are derived from "glove vectors" as can be seen below. Each unique word in the vocabulary corresponds to a specific vector, and these embeddings are preloaded into the model for training. Since raw text data cannot be directly fed into the model, a pre-trained glove model is utilized to obtain these embeddings. An embedding output function is employed to link words to their respective embeddings, allowing for the integration of these representations into the model.

After that authors get embedded metrics to be used for generating word embeddings based on pre-trained word vectors and after shaping it, it gives the value of (1848, 50) which indicates that it contains 1848 words (or vocabulary size) each represented as a 50-dimensional vector.

F. Model Architecture

The model architecture comprises two main components: an image model and a caption model.

The total number of parameters in the model is 1,472,040. These parameters are the weights and biases that the model learns during training. This model appears to be designed for some kind of sequence-to-sequence or caption generation task, where it takes both image features and caption data as inputs and generates an output sequence.

1) Image model

Using "functional API" in Keras for creating a merge model for processing image data. Training the model at the preprocessed data using appropriate loss functions and optimization algorithms. The intention is to minimize the loss and first-rate-song the model's weights.

2) Caption Model

The caption model is based on a sequence model that deals with partial caption sequences. The model is tasked with generating captions that correspond to the given images. The training process involves the utilization of categorical cross-entropy as the loss function, and an appropriate optimization algorithm is employed to update the model's weights iteratively.

G. Training

The training phase involves the amalgamation of the image and caption models. The objective is to train this combined model using preprocessed data. During training, appropriate loss functions and optimization algorithms are employed to minimize the loss and optimize the model's parameters, ultimately enhancing its ability to generate accurate captions for images.

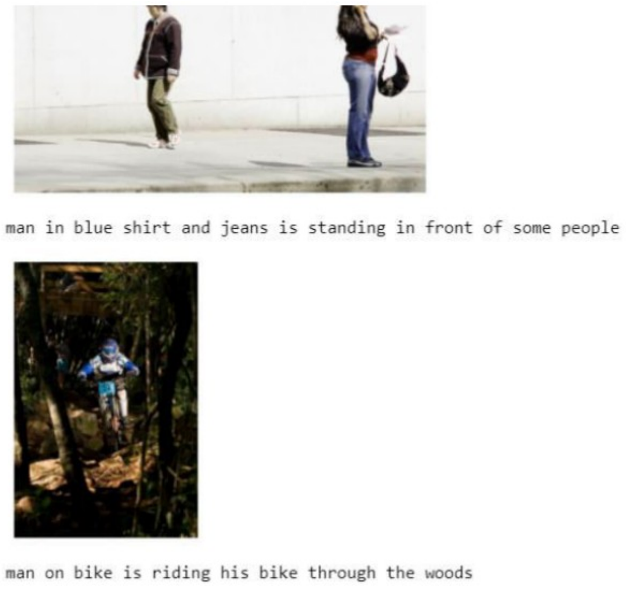
Authors used 10 epochs and the number of pictures used was 3; then a data generate function is called, it will provide output of data generated. The speed of generation will depend on the graphic card, especially, CPU will take a while. Authors utilized CPU in our case, but mostly people will prefer GPU over CPU.

1) Predictions

After creating a function for caption prediction, image's features vector provided, it gave probability of any word in vocab. It will choose the word with maximum probability using "argmax'', because prediction is in the form of indexes, so it is necessary to convert to words, then add to "in_text".

2) Results

By choosing a randomly set which includes 20 images, consequently, authors found out our proposed model is predicting well as can be seen in figure 2.



```
man in blue shirt and jeans is standing in front of some people
```

```
man on bike is riding his bike through the woods
```

Fig. 2. Captions predicted successfully

H. Evaluation

To compare image captions generated through a model, authors have utilized the usage of numerous BLEU (Bilingual Evaluation Understudy) metrics.

Details of the assessment system little by little are broken as follows.

1) Importing NLTK and Calculating BLEU Score

First step is to import the NLTK library. The BLEU score measures how similar the hypothesis sentence is to the reference sentence in terms of n-grams (contiguous sequences of words).

2) Averaging BLEU Scores Over 5

It iterates through the five references for the image, calculates the BLEU score for each reference compared to the same hypothesis, and accumulates these scores. Finally, it prints the average BLEU score for these references.

3) Generating Random Images and Evaluating Captions

This part of the code generates random images (`img_name`) and displays them using Matplotlib. It then generates a caption for each image using a function called `predict_caption`. After generating captions, it proceeds to calculate BLEU scores for each generated caption against the five reference captions (for the same image) using different n-gram weights and prints the results.

4) BLEU Score Calculation with Different Weights

It calculates BLEU scores for the generated hypothesis compared to the five reference captions for the same image using different n-gram weights.

IV. Results and Discussion

The effects of which evaluate the overall performance of a caption generated version by the use of the BLEU Score metric, it offers precious insights of generated captions. A higher BLEU Score suggests higher quality and alignment with human-generated captions.

Table 1: Bleu Score Of 5 Images From Our Proposed Study

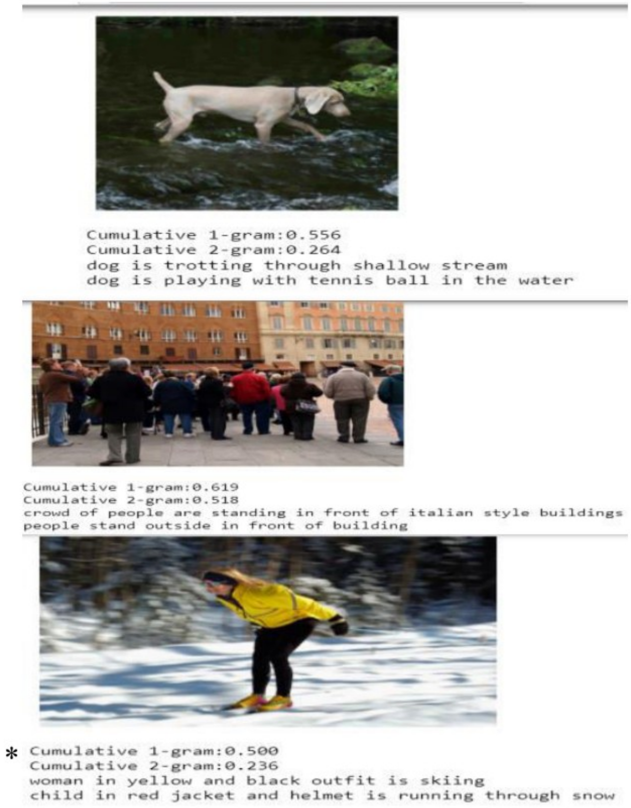
Image Id	Bleu score of 5 images from our proposed study		
	Generated Caption	*BLEU SC-1*	*BLEU SC-2*
Image 1	1. Man holds flag next to snowbound campsite 2. Man in red and white coat is standing in the snow	0.63	0.437
Image 2	1. Blonde haired man is doing demonstration in front of small crowd of people 2. Man with woman sit on the subway	0.143	0.000
Image 3	1. Man in fancy clothing plays guitar on stage 2. Man in red shirt and white shirt is standing in crowd of people	0.231	0.139
Image 4	1. Black and white dog is jumping in the snow at park 2. Two dogs are running through snow covered field	0.250	0.00
Image 5	1. Black and white dog chases pink frisbee 2. Dog is running on the grass	0.564	0.437

Table 1 presents a comparison of image captions generated by a natural language processing model for five different images. The table includes the following columns:

- Image Id: This column identifies each specific image that the generated captions describe.
- Generated Caption: This column contains the textual descriptions of the images generated by the NLP model.
- BLEU SC-1: BLEU SC-1 represents the BLEU score of the generated caption compared to a single reference caption. A higher BLEU score indicates a closer match between the generated caption and the reference caption in terms of n-grams (word sequences) and their frequencies.
- BLEU SC-2: Similar to BLEU SC-1, BLEU SC-2 represents the BLEU score of the generated caption, but in this case, it is compared to two reference captions. This metric provides a different perspective on the quality of the generated text.

Figure 3 are the results of image captioning for each 3 pictures. In our study, authors accomplished a noticeably excessive BLEU Score, indicating that the version has the capability to generate captions which might be linguistically in the direction of human-written captions. This suggests that our model has learned to capture the essence of the pictures and describe them efficiently in textual content. However, it is crucial to notice that whilst BLEU Score gives a beneficial quantitative measure of overall performance, it does not capture all aspects of caption fine. It in general focuses on n-gram overlap and linguistic similarity and won't completely seize the richness and creativity of human-generated captions.

For model architecture, authors have used the combination of an image processing model (CNN) and a sequence model (RNN or Transformer). BLEU Score Performance presented in figure 3 above. BLEU ratings range from 0 to at



```
Cumulative 1-gram:0.556
Cumulative 2-gram:0.264
dog is trotting through shallow stream
dog is playing with tennis ball in the water
```

```
Cumulative 1-gram:0.619
Cumulative 2-gram:0.518
crowd of people are standing in front of italian style buildings
people stand outside in front of building
```

```
* Cumulative 1-gram:0.500
  Cumulative 2-gram:0.236
  woman in yellow and black outfit is skiing
  child in red jacket and helmet is running through snow
```

Fig. 3. Result of image predicted and BLEU Score performance

least one, with higher values indicating better high-quality captions. A perfect match with the reference captions outcomes in a BLEU score of one, even as no overlap effects in a BLEU rating of 0. Typically, better BLEU scores are preferred, but the unique threshold for what constitutes a "desirable" score can range relying on the dataset and studies context.

1) Limitations and Challenges of the usage of TensorFlow and Keras

The challenges of the usage of TensorFlow and Our take a look at utilized TensorFlow and Keras because of the deep getting to know framework and library, respectively, for constructing and educating the caption era version. While that equipment is effective and widely used within the device learning community, they arrive with their very own set of boundaries and challenges.

- *Hardware Dependency*: Training large models with TensorFlow and Keras can be computationally intensive, requiring powerful GPUs or TPUs. This hardware dependency can be a limitation for users without access to such resources.
- *Version Compatibility*: Compatibility issues between different versions of TensorFlow, Keras, and other related libraries could be a challenge when trying to use pre-existing code or models.
- *Community Support*: While TensorFlow and Keras have large and active communities, the rapid pace of development can sometimes lead to a lack of up-to-date documentation or community support for specific issues.

2) Future Directions

Future studies should identify the ability of enhancements in caption generation and address numerous demanding situations. It could indicate satisfactory-tuning for particular domain names, which might include medical imaging, robotics, or artwork, to create more correct and contextually applicable captions. Multimodal techniques, combining textual content and imaginative and prescient, can enhance caption generation by way of incorporating seen information from pix and films. Ethical issues also are crucial as AI-generated content, researchers should recognize mitigating biases, ensure equity, and accountable AI use. Future guidelines must recognize area-specific best-tuning, multimodal procedures, and moral considerations to beautify caption generation and make it greater study, inclusive, and accountable.

3) Analysis

A comparative analysis of similar research:

Research 1 is from [13] paper, second is from [20], and the third is from [12]

Table 2: A Comparative Analysis of 3 Similar Researches

Elements	**Comparative analysis of 3 similar researches**			
	Our research	*Research 1*	*Research 2*	*Research 3*
Main focus	Image captioning	Image captioning	Image captioning	Image caption
Methodology	BLEU Score results for Image caption.	Generating textual descriptions for images using deep learning	Deep learning and NLP for generating image descriptions	Deep learning using CNNs and RNNs
Key technique	BLEU score metrics	Encoder-Decoder	Attention model, CNN, RNN (LSTM)	CNN and RNN
Model architecture	TensorFlow and Keras	Image encoder: CNN, Decoder: RNN	CNN (VGG16 + RNN(LSTM)	CNN(VGG-19), and (LSTM)
Key metrics	BLEU score	BLEU, METEOR, CIDEr, and ROUGE score	BLEU score	NLP-related metrics (BLEU, METEOR, and CIDEr)

These research's results in table 2 all relate to image captioning, with the first three focusing on the development of image captioning models and their respective architectures and techniques. Our study result evaluates the quality of generated captions using the BLEU Score metric and highlights the importance of assessing caption quality beyond linguistic overlap.

V. Conclusion

In this research, performance is a comprehensive evaluation of a caption era model using the BLEU Score metric, offering treasured insights into the pleasantness of generated captions for numerous photographs. Our model exhibited steady and noteworthy overall performance across multiple pix, as evidenced via the high BLEU scores recorded in Table 1. These ratings, starting from 0 to 1, indicate the model's potential to supply captions intently aligned with human-written references. The effects underline the version's talent in know-how photograph content material, as it generated linguistically coherent and contextually applicable captions. Despite the quantitative achievement indicated by means of BLEU rankings, it's essential to know the limitations of this metric in taking pictures of the whole spectrum of caption fine. As highlighted in our discussion, the qualitative richness and creativity inherent in human-generated captions are elements that warrant further exploration and refinement.

Additionally, this takes a look at shedding light on the ethical considerations surrounding AI-generated content. As those technologies emerge as extra established, it's far vital to cognizance of mitigating biases, ensuring fairness, and promoting responsible AI use. Addressing those moral worries is vital for reinforcing the inclusivity and respectfulness of AI-generated captions. Furthermore, the study diagnosed demanding situations associated with the tools used, which include TensorFlow and Keras, emphasizing the need for non-stop improvements in supporting technology for AI research.

In end, whilst our research showcased the version's quantitative prowess thru BLEU scores, destiny efforts should focus on refining the qualitative dimensions of generated captions. By addressing ethical concerns, improving creativity, and improving linguistic richness, AI-pushed technologies can be harnessed efficiently, leading to a destiny in which human-AI collaboration is not most effectively progressive however additionally socially and ethically accountable.

References

[1] "Overview of Image Caption Generators and Its Applications | SpringerLink." Accessed: Oct. 05, 2023. [Online]. Available: https://link.springer.com/chapter/10.1007/978-981-19-0863-7_8

[2] MD. Z. Hossain, F. Sohel, M. F. Shiratuddin, and H. Laga, "A Comprehensive Survey of Deep Learning for Image Captioning," *ACM Comput. Surv.*, vol. 51, no. 6, p. 118:1-118:36, Tháng Hai 2019, doi: 10.1145/3295748.

[3] J.-H. Huang, T.-W. Wu, and M. Worring, "Contextualized Keyword Representations for Multi-modal Retinal Image Captioning," in *Proceedings of the 2021 International Conference on Multimedia Retrieval*, in ICMR '21. New York, NY, USA: Association for Computing Machinery, Tháng Chín 2021, pp. 645–652. doi: 10.1145/3460426.3463667.

[4] S. Mishra, C. S. Minh, H. Thi Chuc, T. V. Long, and T. T. Nguyen, "Automated Robot (Car) using Artificial Intelligence," in *2021 International Seminar on Machine Learning, Optimization, and Data Science (ISMODE)*, Jan. 2022, pp. 319–324. doi: 10.1109/ISMODE53584.2022.9743130.

[5] "SATMeas - Object Detection and Measurement: Canny Edge Detection Algorithm | SpringerLink." Accessed: Apr. 19, 2023. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-031-23504-7_7

[6] "Integrating State-of-the-Art Face Recognition and Anti-Spoofing Techniques into Enterprise Information Systems | SpringerLink." Accessed: Oct. 05, 2023. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-031-45140-9_7

[7] "Image Captioning for Information Generation | IEEE Conference Publication | IEEE Xplore." Accessed: Oct. 03, 2023. [Online]. Available: https://ieeexplore.ieee.org/document/10128347

[8] D. Beddiar, M. Oussalah, and S. Tapio, "Explainability for Medical Image Captioning," in *2022 Eleventh International Conference on Image Processing Theory, Tools and Applications (IPTA)*, Apr. 2022, pp. 1–6. doi: 10.1109/IPTA54936.2022.9784146.

[9] N. Wang *et al.*, "Efficient Image Captioning for Edge Devices." arXiv, Dec. 17, 2022. doi: 10.48550/arXiv.2212.08985.

[10] V. Atliha and D. Šešok, "Image-Captioning Model Compression," *Appl. Sci.*, vol. 12, no. 3, Art. no. 3, Jan. 2022, doi: 10.3390/app12031638.

[11] "Image Captioning Using Deep Learning | IEEE Conference Publication | IEEE Xplore." Accessed: Oct. 03, 2023. [Online]. Available: https://ieeexplore.ieee.org/document/9740788

[12] S. Chakraborty, "Captioning Image Using Deep Learning: A Novel Approach," *Int. J. Res. Appl. Sci. Eng. Technol.*, vol. 11, no. 6, pp. 3468–3472, Jun. 2023, doi: 10.22214/ijraset.2023.54297.

[13] A. Sen, "Captioning Image Using Deep Learning Approach," *Int. J. Res. Appl. Sci. Eng. Technol.*, vol. 11, no. 5, pp. 7425–7428, May 2023, doi: 10.22214/ijraset.2023.53389.

[14] Channasandra, Bangalore, India., N. R. U. S, M. R, and Professor, Department of Computer Science and Engineering RNS Institute of Technology, "IMAGE CAPTIONING: NOW EASILY DONE BY USING DEEP LEARNING MODELS," *Int. J. Comput. Algorithm*, vol. 12, no. 1, Jun. 2023, doi: 10.20894/IJCOA.101.012.001.001.

[15] N. Goel, A. Arora, P. Kashyap, and S. Varshney, "An Analysis of Image Captioning Models using Deep Learning," in *2023 International Conference on Disruptive Technologies (ICDT)*, May 2023, pp. 131–136. doi: 10.1109/ICDT57929.2023.10151421.

[16] "Deep Image Captioning: An Overview | IEEE Conference Publication | IEEE Xplore." Accessed: Oct. 03, 2023. [Online]. Available: https://ieeexplore.ieee.org/document/8756821

[17] "[1906.05963] Image Captioning: Transforming Objects into Words." Accessed: Oct. 03, 2023. [Online]. Available: https://arxiv.org/abs/1906.05963

[18] J. Pavlopoulos, V. Kougia, and I. Androutsopoulos, "A Survey on Biomedical Image Captioning," in *Proceedings of the Second Workshop on Shortcomings in Vision and Language*, Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 26–36. doi: 10.18653/v1/W19-1803.

[19] "[1905.13302] A Survey on Biomedical Image Captioning." Accessed: Oct. 03, 2023. [Online]. Available: https://arxiv.org/abs/1905.13302

[20] L. Panigrahi, R. R. Panigrahi, and S. K. Chandra, "Hybrid Image Captioning Model," in *2022 OPJU International Technology Conference on Emerging Technologies for Sustainable Development (OTCON)*, Feb. 2023, pp. 1–6. doi: 10.1109/OTCON56053.2023.10113957.