Image Captioning of an Environment Using Machine Learning Algorithms (A Case Study of Gwarzo Road, Kano Nigeria)

Muhammad Aliyu[1], Amir Abdullahi Bature[2]

[1]Research Scholar, Bayero University, Kano (Nigeria)

[2]Associate Proffessor, Bayero University, Kano (Nigeria)

DOI: https://doi.org/10.51244/IJRSI.2024.1110055

Received: 11 October 2024; Revised: 20 October 2024; Accepted: 24 October 2024; Published: 20 November 2024

ABSTRACT

This paper investigates the application of machine learning algorithms for automatic image captioning, focusing on a case study of Gwarzo Road in Kano, Nigeria. The research aims to design a robust VGG16/LSTM-based model that generates accurate and contextually relevant descriptions for images captured along the Kabuga to Bayero University Kano new site route. The methodology involves collecting images at three distinct times of the day (morning, afternoon, and evening) over 60 days, resizing and labelling them with relevant captions to build a comprehensive dataset. The VGG16 model, known for its efficiency in image processing, was employed for feature extraction, while the LSTM network was used to generate captions by interpreting the contextual and semantic details of the images. This study addresses key challenges in image captioning, such as localized object detection and generating meaningful textual descriptions, improving on existing datasets and models that often lack contextual relevance in specific environments. The expected outcomes of this research include the development of a precise caption generation model with high accuracy and efficiency. The resulting model achieved a BLEU score of 0.051, representing baseline performance in caption generation with partial alignment to human-generated references. Additionally, the model's highest accuracy based on the loss function reached 55%, while the lowest accuracy was 50%, with an average accuracy of 53%. The creation of a localized image database further enhances the significance of this research for future applications and studies in image captioning.

Keywords: Image Captioning, VGG16, LSTM, Machine Learning, Object Detection, BLEU Score, Kano, Gwarzo Road, Localization, Image Database.

INTRODUCTION

Computer vision is a rapidly evolving field that aims to equip artificial agents with the ability to perceive and interpret complex visual scenes, much like human beings. As a core area of artificial intelligence (AI) and machine learning (ML), computer vision has been a subject of intense research for decades. Despite significant advancements, human vision systems still outperform artificial ones in most tasks. Another crucial aspect of human interaction is language—our primary means of communication. Developing an AI agent capable of communicating through language is a significant goal for improving human-agent interactions and leveraging the vast amount of knowledge available in human language.

Natural Language Processing (NLP), a subfield of AI and ML, aims to enable machines to understand, interpret, and generate human language. Although NLP has seen substantial progress, many challenges remain, particularly when integrated with computer vision. The task of image captioning, where a system generates textual descriptions for images, lies at the intersection of these two domains computer vision and NLP. This task requires the model to both visually comprehend an image and translate its understanding into

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XI Issue X October 2024

coherent, contextually accurate language (Abhijit Roy, 2020). The goal of image captioning is to teach a computer to interpret images and describe them in natural language. Achieving this requires a deep understanding of both the visual content and the semantics associated with the objects and relationships within the image. To generate accurate captions, a system must not only recognize the objects but also understand their interactions and context within the scene. This process, known as *feature extraction*, plays a critical role, as it gathers detailed information about the image's components and their significance. Image captioning is fundamental in the deep learning domain because it combines both image processing and language generation to create meaningful descriptions. This task has seen significant advancements due to the development of neural network models such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs). CNNs are widely used for visual feature extraction, while RNNs are employed to generate the corresponding captions by understanding the sequential nature of language (Zhishan et al., 2020).

Image captioning is not only vital for practical applications, such as automatic image indexing and description generation, but also serves as a benchmark for image understanding, including object recognition, scene comprehension, and the detection of object-object relationships. In recent years, the challenge of image captioning has gained attention due to advancements in deep learning, the availability of large datasets, and increased computational power. Modern models can now generate captions for images, achieving results that were once deemed impossible (Megha et al., 2021). The ability to automatically describe the content of images using natural language has far-reaching implications. For example, this technology could assist visually impaired individuals by providing detailed descriptions of online images, or enhance the accuracy of image and video analysis in scenarios such as social media sharing and surveillance systems (Jianhui Chen et al., 2019). This research paper aims to contribute to these advancements by developing a VGG16/LSTM model to generate captions for images taken along Gwarzo Road in Kano, Nigeria. To achieve the aim, the following objectives were followed. Developed a database of images From Kabuga area to Bayero University Kano new site, labeled all the images with appropriate captions manually, developed efficient AI model for images caption using combined VGG16 and LSTM and validate the performance of the developed model.

LITERATURE REVIEW

D. Bahdanau et al. (2014) explored retrieval-based captioning, which was a widely used method for generating captions by retrieving the most relevant text from an existing caption pool for a given query image. The retrieved caption could either be directly taken or composed from various components. Their method involved identifying a visually similar image from the training dataset and used the Lin similarity measure to assess the semantic distance between images, aided by parsing techniques like the Curran and Clark parser. K. Cho et al. (2014) pointed out limitations in retrieval-based image captioning, such as the inability to describe new object combinations. Although such methods produced grammatically correct captions, the descriptions often failed to match the actual scene, limiting their utility in dynamically generating relevant captions for images with novel content. Karpathy et al. (2015) suggested an improvement by ranking captions based on a probability density function for words, retrieving the highest-scoring captions for query images. They highlighted the effectiveness of combining retrieved phrases for query image description, using Stanford Core NLP to match relevant sentences from a dataset and providing a global view of image features.

Qi Wu et al. (2018) introduced a method to better correlate training images and their captions using Kernel Canonical Correlation, which mapped images and text into a shared space to compute cosine similarities for better caption retrieval. This method attempted to reduce noise in visual data when captioning images. Simao Herded et al. (2019) proposed the "Object Relation Transformer" model, which gave special attention to the geometric relationships between recognized objects in images. Their encoder-decoder architecture used object detector region recommendations to produce improved image captions, showing enhanced performance on the MS-COCO dataset. A. Oluwasanni et al. (2019) introduced Fully Convolutional

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XI Issue X October 2024

CaptionNet (FCC) for describing differences between images. FCC employed an encoder-decoder framework for extracting visual features and computing feature distances to generate captions. Their experiments demonstrated that FCC outperformed existing models on benchmark datasets by generating concise and meaningful textual differences. Guanghui Xu et al. (2021) addressed the challenge of generating multiple captions to describe different parts of an image.

Their approach tackled difficulties like selecting which part of an image to describe and managing the complex relationships between objects, achieving state-of-the-art performance through their multiple caption generation method. Jing Wan et al. (2021) enhanced OCR-based image captioning by focusing on the geometric relationship between OCR tokens in images. Their proposed LSTM-R network integrated the learned relationships with visual and semantic representations, resulting in more accurate descriptions of images that contain OCR tokens. Shitiz Gupta et al. (2021) developed a novel architecture using transfer learning and Stacked LSTMs with soft attention to improve image captioning accuracy. Their model utilized several transfer learning techniques and achieved higher accuracy in image captioning, as validated by benchmark datasets and metrics like BLEU and METEOR. Megha J et al. (2021) designed a deep learning model to detect objects and their relationships in images to generate captions. They used the Flickr8k dataset and employed transfer learning with the Xception model, showing that such methods could aid in applications like image segmentation and assisting visually impaired individuals.

Sulabh & Samir (2022) presented a streamlined version of SqueezeNet combined with the RMS Prop algorithm for image captioning. Their system, which focused on visual information extraction and caption generation, achieved an accuracy of 93.46%, demonstrating the effectiveness of their hyper-tuned model. Theyi Qi Yan and Yulin Zhu (2022) discussed the use of spatiotemporal changes and template fulfilment for generating storylines from images, contrasting it with traditional image captioning methods. Their approach combined pattern recognition with deep learning to better couple contextual information with images, resulting in more detailed descriptions. M. Hartmann et al. (2022) proposed an interactive learning approach for image captioning, aiming to optimize human feedback through data augmentation. This method reduced the need for large amounts of supervised data, making human feedback more efficient for training image captioning models. Antonio M et al. (2023) combined multiple deep neural networks for hierarchical object detection, improving caption generation through a natural language processing module. Their framework demonstrated a significant improvement in captioning accuracy, outperforming single models. Md Adnan Wasi et al. (2023) focused on deep learning techniques for generating descriptive captions using CNNs for feature extraction and RNNs for sequential language generation. Their paper contributed practical insights into advancements in computer vision and NLP by developing a functional image captioning system. Almost all the authors in the review papers focused on CNN and RNN, which limitations include less number of layers (neuron) and vanishing of gradients respectively. Our proposed machine learning algorithms VGG16 and LSTM remedy the limitations of conventional algorithms.

RESEARCH METHODOLOGY

Methods of Data Collection

For this study, the proposed model aimed to develop captions for images captured in our local environment of Gwarzo Road, Kano, Nigeria. The project was designed after reviewing similar efforts focused on enhancing the performance of image captioning models in different environments. The images were taken over a period of 60 days, from May 18th to July 20th, 2023, during three distinct sessions of the day:

- Morning (7:00 am to 10:00 am)
- Afternoon (12:00 pm to 3:00 pm)
- Evening (4:00 pm to 7:00 pm)

This time-based division allowed for the collection of images with varying light conditions, helping to create a more diverse dataset. The image collection was carried out alongside a professional photographer. **Keke

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XI Issue X October 2024

Napep** (tricycle) and **motor vehicles** were used as modes of transportation during the sessions to snap images from different locations along the route. The collected images were then resized to **224x224 pixels** for uniform processing and compatibility with the model. Each image was labelled manually with appropriate captions as a human reference to aid in training and validating the model. The image captioning model was developed using a pre-trained VGG16 algorithm, which was employed to process the images and extract relevant features. For the text data, a **Long Short-Term Memory (LSTM)** algorithm was used to train the textual information and classify it, ultimately generating suitable captions for the images. The combination of these two algorithms—VGG16 for visual feature extraction and LSTM for text generation—forms the backbone of the proposed model, enhancing its ability to generate accurate captions.

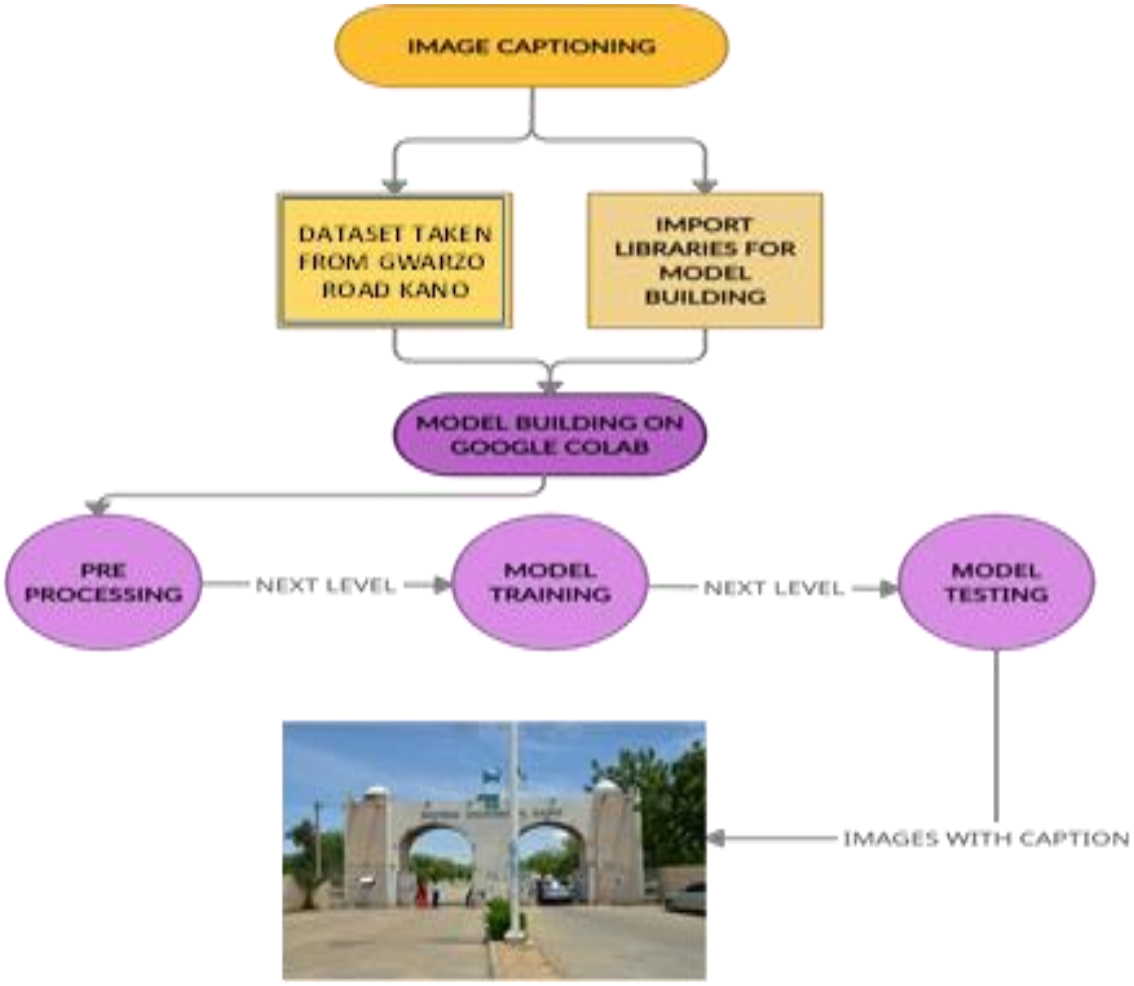


Figure 1: Model development approach

Materials and Software Libraries Used

The collected images were processed, trained, and tested using **Google Colaboratory**, a cloud-based platform that provides the necessary computational resources for running machine learning models. The dataset containing the images was loaded into the Colaboratory environment, and the file path for the dataset was specified accordingly to ensure smooth access during both training and testing phases.

To develop the model, the following software libraries were imported:

1. **TensorFlow**: Used as the primary framework for building and training the deep learning model. TensorFlow provided tools to implement both the image processing using the VGG16 model and the text generation using LSTM.
2. **Pandas**: This library was used for data manipulation, including loading and organizing image metadata and labels into a structured format for further processing.

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XI Issue X October 2024

3. **NumPy**: Essential for performing numerical operations, NumPy was utilized for handling array data and performing mathematical operations required during model training.
4. **Matplotlib**: Employed to visualize the results and training progress, Matplotlib helped in plotting learning curves and analysing the performance of the model during various phases of development.

Model Training and Modification of VGG16

To train a deep learning model for image captioning, the first step is to convert the images into features that can be used as input. In this project, the **VGG16** model, a Convolutional Neural Network (CNN), was employed to extract image features due to its high accuracy in image recognition tasks. The VGG16 model has been pre-trained on large image datasets, which makes it highly effective in extracting meaningful features from new images.

Key Features of VGG16

The VGG16 model consists of 16 layers of weights that contribute to its ability to extract detailed visual features from images. It utilizes **3x3 convolutional layers** followed by **max-pooling** layers to compress images while retaining critical features.

By removing the final classification layer, which typically predicts image classes, the internal feature representation of the image is captured just before the classification stage. This internal representation is used for further processing.

- **Feature Extraction**

In this process, the images are fed into the VGG16 model to obtain a set of feature maps representing the internal characteristics of each image. These features serve as the input for further captioning tasks, and the quality of feature extraction is crucial to the overall performance of the model.

- **Image Captioning**

The **picture captioning task** requires combining the visual features extracted from the images with textual information. Once the features are extracted, they are used as input to train the model in conjunction with text data. The goal of this process is to generate captions that accurately describe the content of the images. Research indicates that increasing the number of layers in a network, such as in the VGG16 model, improves the quality of visual feature extraction, which in turn enhances the performance of image captioning.

- **Customized Dataset**

Two key text files, **Customized dataset.trainImages.txt** and **Customized dataset.devImages.txt**, were used to manage the training and development data. These files contain the image IDs that correspond to the images used in the training and development phases. They allow the model to identify the images and their respective captions from the dataset.

- **Text Data Encoding**

In addition to image features, the model requires the caption text to be transformed into a numerical format for input. Each word in the caption is encoded with a numerical value, and these encoded words are then fed into the model. The model processes the numerical features extracted from both the images and the captions to perform mathematical operations required for machine learning. The diagram below shows the flow chart of the developed model.

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XI Issue X October 2024

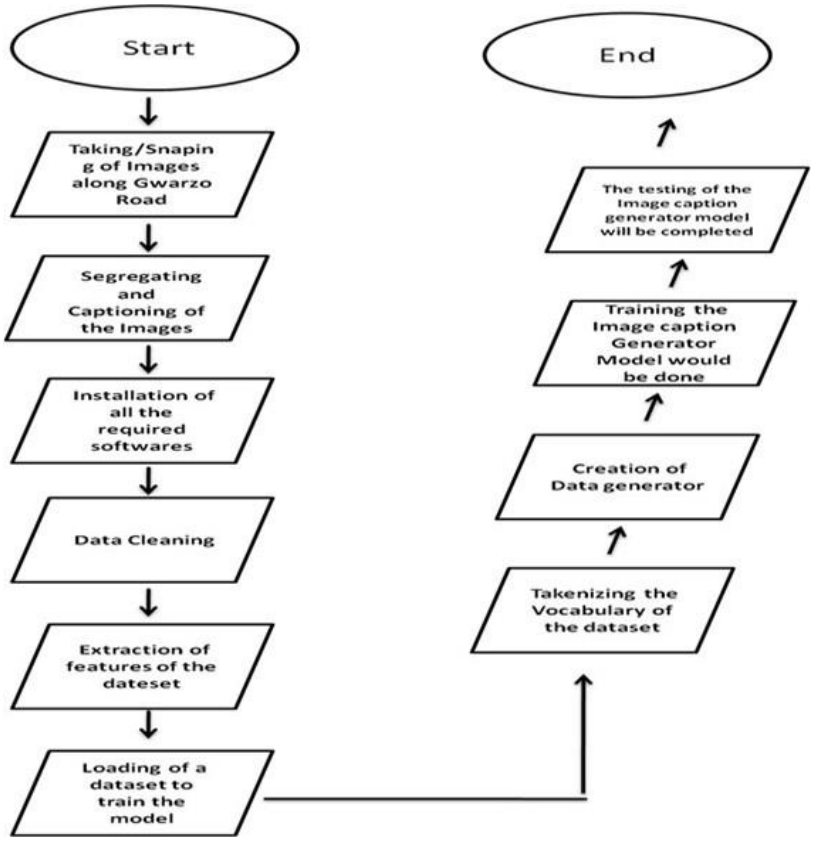


Figure 2: Flow chart of the model

Dataset

The dataset utilized for this project comprises both images and corresponding text descriptions, specifically curated for the task of image captioning. The images were captured in real-world scenes and depict objects and activities around **Gwarzo Road** in **Kano State**. Each image is accompanied by a caption that describes the content of the image in detail, allowing the model to learn the relationship between visual elements and their textual descriptions.

Dataset Composition

Total 200 images, each with a corresponding caption. Image Source were taken from various scenes, including actions and objects in the Gwarzo Road area. Text Descriptions a text file accompanies the dataset, containing captions for each image. These captions describe the scene or object in the corresponding image and are used as ground truth for training the model.

Dataset Structure:

The dataset is divided into separate folders for images and text:

- Image Folder: Contains all the images with unique IDs assigned to each image. The unique ID helps map each image to its respective caption.
- Text Folder: Contains the captions, each linked to its corresponding image by the unique ID.

Table no 1: Dataset Description

Dataset Name	**Size**		
	Train	**Valid**	**Test**
Gwarzo_Buk	**100**	**50**	**50**

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XI Issue X October 2024

RESULT AND DISCUSSION

The figure below shows a summary of a neural network model used for training and evaluating the model.

Model: "model_4"

Layer (type)	Output Shape	Param #	Connected to
input_7 (InputLayer)	[(None, 4096)]	0	[]
dense_6 (Dense)	(None, 256)	1048832	['input_7[0][0]']
input_8 (InputLayer)	[(None, 16)]	0	[]
reshape_2 (Reshape)	(None, 1, 256)	0	['dense_6[0][0]']
embedding_2 (Embedding)	(None, 16, 256)	88320	['input_8[0][0]']
concatenate_2 (Concatenate)	(None, 17, 256)	0	['reshape_2[0][0]', 'embedding_2[0][0]']
lstm_2 (LSTM)	(None, 256)	525312	['concatenate_2[0][0]']
dropout_4 (Dropout)	(None, 256)	0	['lstm_2[0][0]']
add_2 (Add)	(None, 256)	0	['dropout_4[0][0]', 'dense_6[0][0]']
dense_7 (Dense)	(None, 128)	32896	['add_2[0][0]']
dropout_5 (Dropout)	(None, 128)	0	['dense_7[0][0]']
dense_8 (Dense)	(None, 345)	44505	['dropout_5[0][0]']

Total params: 1739865 (6.64 MB)
Trainable params: 1739865 (6.64 MB)
Non-trainable params: 0 (0.00 Byte)

Figure 3: Training of model using Google Collaboratory.

Below are some of the images used to test the trained model, the trained model created the descriptions for the pictures as shown below

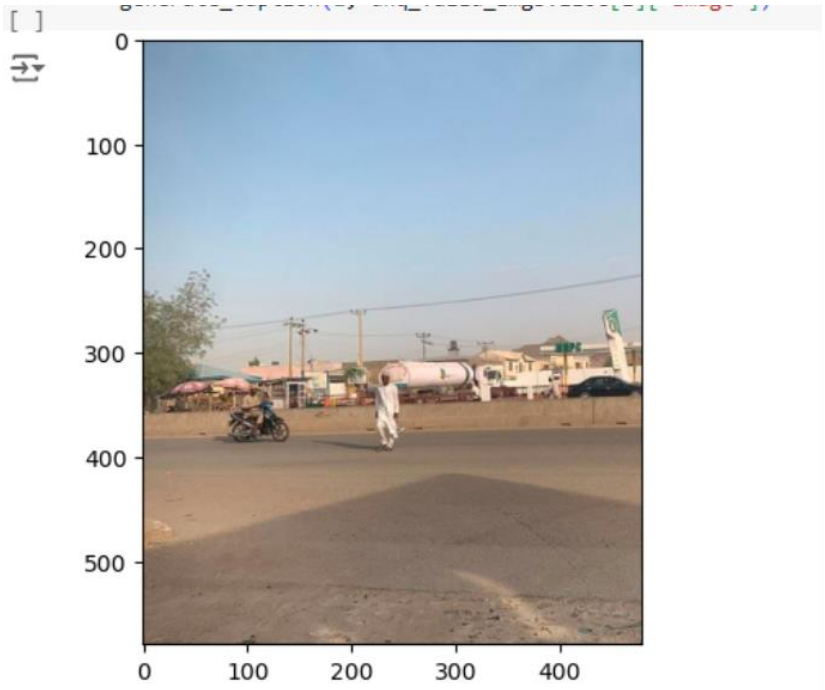


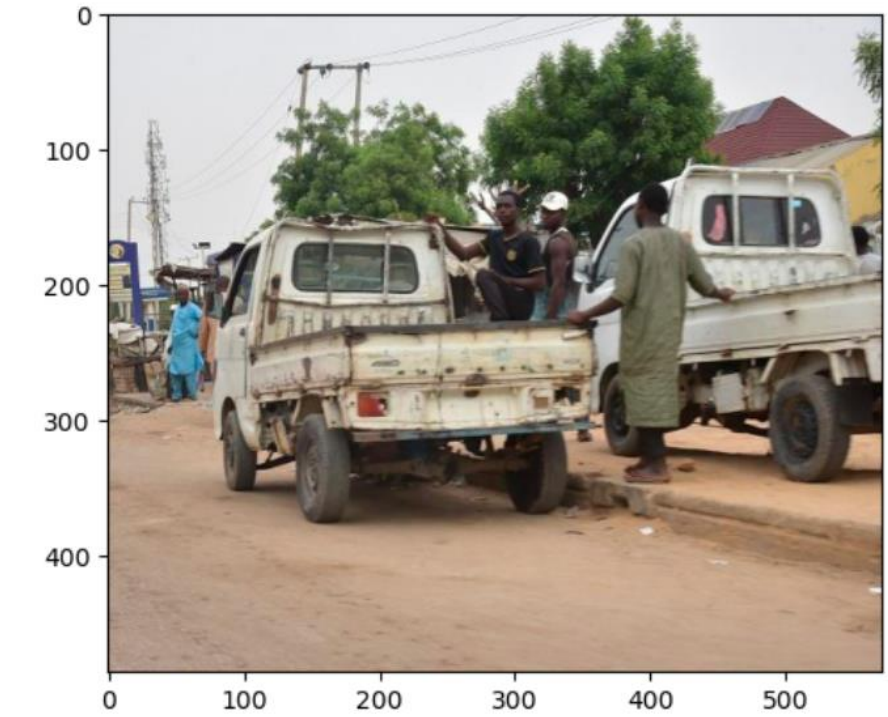
Figure 4: A man with white dress crossing the road.

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XI Issue X October 2024


```
Actual Caption :
['A Boy With T-Shirt Holding a Sachet of Water']


Predicted caption :
man with white shirt in front of conveying.
```

Figure 5: A boy with T-shirt holding a sachet water.




```
Actual Caption :
['A White Mini Pick Up Truck']


Predicted caption :
some truck on the road.
```

Figure 6: A white mini Pick Up truck.

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XI Issue X October 2024

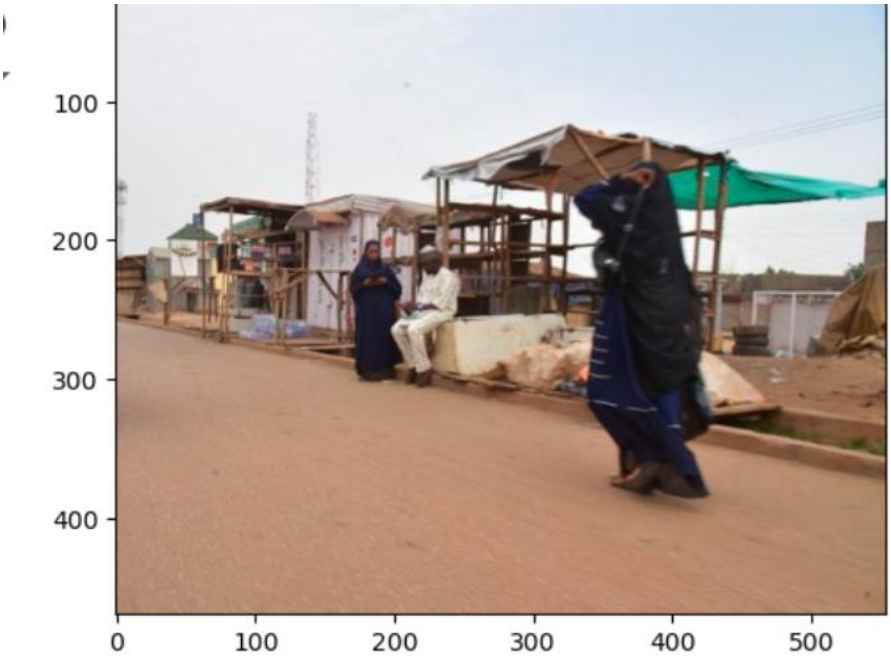


Figure 7: A man sitting and a woman standing waiting for a ride.

Table 2 summarizes the accuracy percentages achieved by the model at each epoch, alongside the corresponding loss function values. The results indicate fluctuations in accuracy, with a noticeable peak at epoch 7, where the accuracy reached 100%. However, this may suggest overfitting, as subsequent epochs show a decline in performance.

Table 2: Accuracy percentage based on minimal loss function

EPOCHS NUMBER	LOSS FUNCTION	ACCURACY (%)
1	4.0393	89.82
2	4.1346	87.75
3	4.0883	88.75
4	4.0043	90.61
5	4.0087	90.51
6	3.9187	92.59
7	3.6282	100.00
8	4.0217	90.23
9	4.5034	80.57
10	3.7651	96.37
11	3.9638	91.53
12	4.0013	90.68

The accuracy fluctuates throughout the epochs, with some epochs showing a high accuracy percentage, particularly around epoch 6 (92.59%) and 10 (96.37%). The model achieved its best accuracy during epoch 7, but the performance declined in subsequent epochs.

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XI Issue X October 2024

Runtime and Loss Function Analysis

Table 3 provides an analysis of accuracy in relation to the time taken for each epoch and the corresponding loss functions. This allows for understanding how computational efficiency correlates with model performance.

Table 3: Accuracy as per run time and loss function

S.NO	LOSS FUNCTION	TIME TAKEN	TIME TAKEN PER STEP	ACCURACY (%)
1.	4.0393	2 sec	40ms/step	51.683%
2.	4.1346	752 ms	40ms/step	51.240%
3.	4.0883	751 ms	49ms/step	50.807%
4.	4.0043	808 ms	48ms/step	55.634%
5	4.0087	822 ms	41ms/step	50.176%
6.	3.9187	865 ms	41ms/step	55.326%
7.	3.6282	878 ms	40ms/step	53.817%
8.	4.0217	772 ms	41ms/step	52.850%
9.	4.5034	789 ms	40ms/step	54.208%
10.	3.7651	807 ms	44ms/step	54.339%
11.	3.9638	834 ms	41ms/step	52.334%
12.	4.0013	780 ms	40ms/step	51.396%

From Table 3, it can be observed that the accuracy percentages do not consistently correlate with the loss function values or the time taken. For instance, while epoch 1 shows a high loss function of 4.0393, its accuracy is significantly lower compared to later epochs. Conversely, epoch 6, with a loss function of 3.9187, shows a relatively high accuracy of 55.326%. The model achieved the highest accuracy (55.634%) at epoch 4, where the loss function was 4.0043, demonstrating a balanced performance relative to training time. The training results indicate that the model has a solid foundation for generating image captions, with fluctuations in performance metrics across epochs. The model appears to overfit after epoch 7, necessitating further tuning and regularization strategies to enhance generalization. Further analyses should explore the impact of learning rates, batch sizes, and data augmentation techniques on model performance.

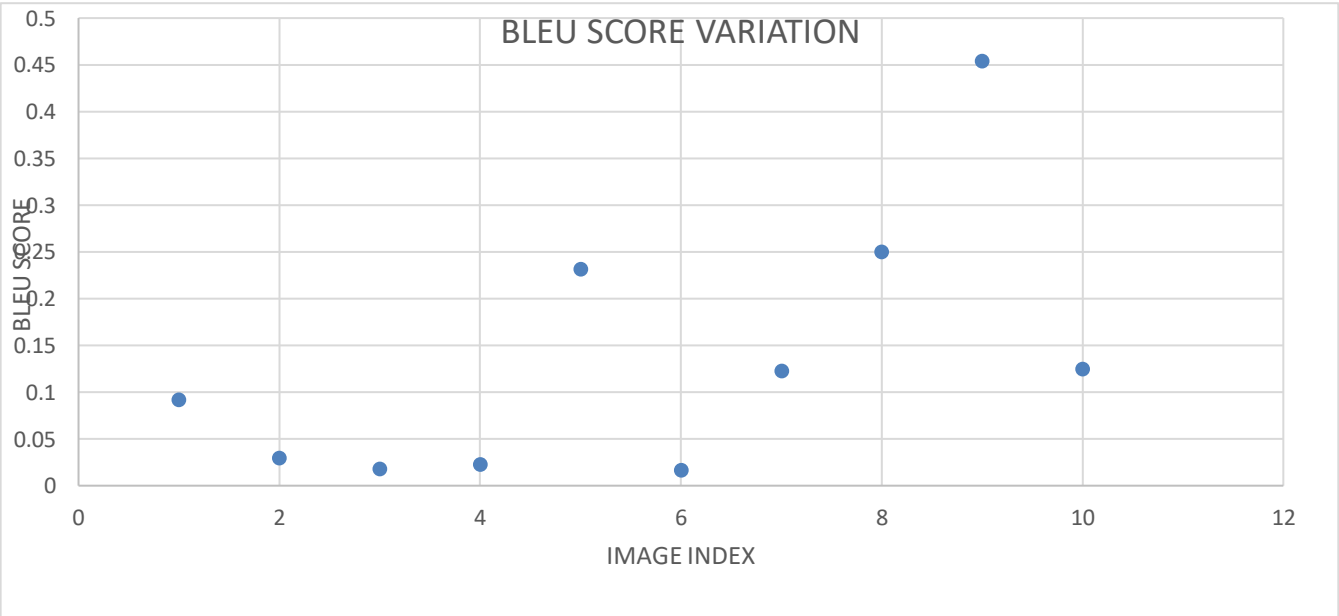


Figure 8: Performance of the proposed model in terms of BLEU Variation

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XI Issue X October 2024

The BLEU (Bilingual Evaluation Understudy) score is a widely used metric for evaluating the quality of text generated by machine learning models, particularly in the context of natural language processing tasks such as machine translation and image captioning. In this section, we analyse the BLEU score variations across different image indices as shown in Figure 1.

Table 3: Evaluation Matrices

NAME	VALUE
BLEU Score	0.051
Highest accuracy	55%
Average Accuracy	53%
Lowest accuracy	50%

The BLEU score of **0.051** indicates a relatively low level of agreement between the generated captions and the reference captions. This score reflects the precision of the model in capturing the essential elements of the images through textual descriptions. Given that the BLEU score is typically normalized, this suggests that there may be significant room for improvement in the model's captioning capabilities. The highest achieved accuracy of **55%** signifies that for a portion of the test dataset, the model was able to generate captions that matched the reference captions satisfactorily. This indicates that some images were processed effectively, though this is still a modest achievement overall. An average accuracy of **53%** demonstrates that across all test instances, the model's performance is inconsistent. While some images received accurate captions, others did not, leading to this average figure. The lowest accuracy recorded at **50%** indicates that there are images for which the model generated captions that were no better than random guessing. This is particularly concerning and points to the necessity for targeted improvements in the model.

CONCLUSION

In this study, we developed an image captioning model that integrates the VGG-16 architecture with Long Short-Term Memory (LSTM) networks. Our primary aim was to create a system capable of generating meaningful captions for images captured in the Gwarzo Road area of Kano State. The methodology involved meticulous data collection, resulting in a dataset of 200 images taken over a period of 60 days, which was systematically divided into training, development, and testing sets. This structured approach aimed to ensure a robust dataset that could effectively train the model, facilitating the use of advanced libraries such as TensorFlow, Pandas, NumPy, and Matplotlib for model training and evaluation. The results of our experiments revealed that the model achieved an average accuracy of 53% and a peak accuracy of 55%, indicating some success in generating relevant captions. However, the BLEU score of 0.051 highlighted limitations in the model's ability to produce captions that closely aligned with human-generated descriptions. This discrepancy suggests that while the model demonstrates foundational capabilities, there are considerable gaps in its performance that need addressing to enhance the quality of generated captions. This work significantly contribute to a database of images taken at Gwarzo road from Kabuga to BUK new site Kano and an image caption generation model with a good degree of precision and less number of epochs.

Looking ahead, several areas warrant further exploration and improvement. Enhancing the dataset's quality and diversity by incorporating a broader range of images could significantly improve the model's contextual understanding and semantic richness. Additionally, refining the VGG-16 architecture and experimenting with alternative models may yield better feature extraction and overall performance. Training strategies, including data augmentation and transfer learning from more extensive and diverse datasets, could also help mitigate current accuracy limitations. This research contributes valuable insights into the field of image captioning, particularly regarding the challenges faced and potential strategies for improvement. While the initial results provide a promising foundation, significant work remains to enhance the model's performance. Future efforts will focus on refining the architecture and methodologies employed to develop a more

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XI Issue X October 2024

sophisticated image captioning system capable of generating high-quality textual descriptions that accurately reflect the content of the images.

REFERENCES

1. **Guanghui Xu et al., (2021),**" Towards accurate text-based image captioning with content diversity exploration," Computer Vision Foundation, IEEE explore.
2. **Jing Wang et al., (2021),"** Improving OCR-based Image Captioning by Incorporating Geometrical Relationship," Computer Vision Foundation, IEEE explore.
3. **Qi Wu, Chunhua Shen, Peng Wang, Anthony Dick, and Anton van den Hengel. (2018).** Image captioning and visual question answering based on attributes and external knowledge. IEEE transactions on pattern analysis and machine intelligence 40, 6, 1367–1381.
4. **Karpathy A, Fei-Fei L (2015),** Deep visual-semantic alignments for generating image descriptions. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3128– 3137.
5. **Fang, Hao, et al. (2015)"**From captions to visual concepts and back." Proceedings of the IEEE conference on computer vision and pattern recognition.
6. **Xu, Kelvin, et al. (2015)"**Show, attend and tell Neural image caption generation with visual attention." International Conference on Machine Learning.
7. **Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. (2015)**. Neural machine translation by jointly learning to align and translate. In International Conference on Learning Representations (ICLR).
8. **Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang**. **(2017).** Bottom-up and top-down attention for image captioning and via. arXiv preprint arXiv:1707.07998.
9. **Mareike Hartmann et al., (2022),"** Interactive Machine Learning for Image Captioning", arXiv.
10. **Mohamed Omriet al., (2022),"** Modeling of Hyperparameter Tuned Deep Learning Model for Automated Image Captioning", Mathematics2022, 10, 288.
11. **Zhishan Yang, Noaki Okazaki (2020)** "Image Caption Generation For News Articles":proceedings of the 28th International conference on computational linguistics, pages 1941-1951 2020.
12. **Megha J Paricker, et al, (2021)** "image caption Generator" International Journal of Innovative Technology and exploring Engineering (IJTEE) ISSN: 2278-3075(online) Volume -10 issue -3 January 2021.
13. **Jianhui Chen, Wenguiag Dong, Micchenli (2019).** "Image Caption Generator based on Deep Neural Network". http://www.researchgate.com/ [online: access 20-December-2022].
14. **Ralf C, Eric Morris.** "Understanding LSTM". A tutorial into LSTM RNN. September 2019.
15. **Shitiz Gupta et al, (2021)** "Image Caption generation and comprehensive comparison of image". Fusion: practice and applications (FPA). Vol No 2 page 42-55. 2021.
16. **Ariyo Oluwasanmi et al, (2019).** "Fully Convolutional Captionet": Siamse difference captioning attention Model. IEEE access multidisplinary 10.1109 aceess 2019.
17. **Abhijit Roy (2020).** "A guide to image captioning"http: //towardsdatascience.com.[online :Access on 20-December-2022].
18. **Vijay Choubey, (2020**). "Understanding RNN and LSTM" http://medium.com/analyics.vidhaya [online: access 20-December-2022].
19. **Rohini G, (2021).** "Everything you need to know about VGG16" http://www.medium .com/mygreat [online: access on 20- December 2022].
20. **Sulabh and Samir (2022)** "Image Captioning using Deep Stacked lstms, Contextual Word Embeddings and Data Augmentation." 1 flickr.com.
21. **Yulin zhu Theyi Qi Yan et , (2022),"** Image-Based Storytelling Using Deep Learning", The 5th International Conference on Controland Computer Vision (ICCCV).
22. **Simao Herded et , (2019)** "Image Captioning: Transforming Objects into Words"https:// github.com/yahoo/object relation transformer

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XI Issue X October 2024

23. **Antonio M. Rinaldi, Cristiano Russo, Cristian Tommasino,** ”Automatic image captioning combining natural language processing and deep neural network” Results in Engineering 18 (2023) 101107. http://www.sciencedirect.com/journal/results-in-engineering [online: access on 24- July-2023].
24. **Md Adnan Wasi, Rakesh Das, Purnendu Sarkar, Suvajit Singha, Tanmay Barman, Sourov Kumar Kundu, Moloy Dhar, Sayan Roy Chaudhuri (2023).”** Image Captioning Using Deep Learning” International Journal for Research in Applied Science & Engineering Technology (IJRASET) ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538 Volume 11 Issue VI Jun 2023- http://www.ijraset.com.