**Sports Image Caption Generator**

**Abstract**

One major difficulty in the field of artificial intelligence (AI) is using automatic image captioning to provide descriptions for photographs that are human-like. This necessitates knowledge of the context in addition to the content, particularly when analyzing broad scenery and figuring out the relationships between objects. This paper offers a new method for captioning sports videos with a deep learning network that mixes sequential modeling (LSTM) networks for long short-term memory and Feature extraction using convolutional neural networks (CNNs). Initially, the model is fine-tuned on a carefully selected sports picture dataset, following training on the Flickr8k dataset. A comprehensive review of the literature on current techniques and datasets for automatic image captioning highlights both state-of-the-art techniques and areas that still need development. To evaluate the performance of the model across various datasets of sports images, the model setup uses CNNs and LSTMs, where the former is used for visual feature extraction and the latter is used for producing logically appropriate captions. The results show that the training technique improves automated descriptions for noteworthy events in sports images greatly in terms of accuracy and relevance. Through the integration of cutting-edge deep learning algorithms with insights from literary studies, sports picture captioning improves the field of automatic image captioning and sets the standard for human-like sports image descriptions.

**Keywords:** CNN, LSTM, image captioning, deep learning.

**Introduction**

A popular dataset for studies on image captioning and visual-semantic alignment tasks is Flickr8k, which is a collection of photographs together with textual descriptions that go with them. It consists of 8,000 images collected from Flickr, together with five reference captions provided by human annotators typically through platforms like Amazon Mechanical Turk as seen in paper titled "Collecting Image Annotations Using Amazon's Mechanical Turk" by Micah Hodosh, Peter Young, and Julia Hockenmaier (2013) [7]. The dataset is also used in tasks related to visual-semantic alignment, where the goal is to associate images with relevant textual information. The images were sourced from Flickr, a popular photo-sharing platform. The dataset was introduced at the University of Illinois at Urbana-Champaign as seen in “Framing Image Description as a Ranking Task: Data, Models and Evaluation Metrics” by Micah Hodosh et.al; (2013) [8]. But for the prescribed model we have chosen a part of the Flickr8k dataset which contains only sports images and is suitable for the given problem statement of sports image captioning.

Automatic image captioning is the process of generating human-like sentences for images through computational methods and it is a large step forward with huge potential and padding on many fields such as industry, security, and medicine: “Convolutional Image Captioning” by Jyoti Aneja et al; (2018) [9] “Medical image captioning via generative pretrained transformers” by Alexander Selivanov et. al; (2023) Automatic Caption Generation for Medical Images Imane Allaouzi et.al; [10] and “Image Captioning using Deep Learning: A Systematic Literature Review” by Murk Chohan et al; (2020) [11]. Automatic image captioning steps into a more complex area compared to conventional computer vision tasks, which are primarily about object recognition. It calls for not only recognizing items inside an image but also recognizing the image's overall contents in a way that mimics how people process visual information, with regards to human perception as seen in “Spatial as Deep: Spatial CNN for Traffic Scene Understanding” by Xingang Pan et al; (2018) [13].

Bringing robots closer to human behavior is the goal of automation and AI research. We go through several procedures to automatically caption photographs. Prior to writing a description of the objects in the picture, we identify them, determine their nature, and consider their connections, as seen in “Automatic Image Captioning Based on ResNet50 and LSTM with Soft Attention" by Yan Chu et al; (2020) [15]. The main focus of object recognition and image classification is the identification of individual items. Recognizing separate elements is the key to classifying photos and identifying items, as in “Automatic image captioning" by P. Duygulu et al (2004) [14], A clear understanding of the images is required to get the best results in comparison to handwritten ones generated by automatic photo description, as seen in “Transforming remote sensing images to textual descriptions" by Usman Zia et al; (2022) [16].

Object detection is necessary for picture image captioning to achieve human-like accuracy. This presents a challenging task: association detection, relationship identification, feature extraction, and text description synthesis. Different deep learning systems were employed by scientists to address these problems. For instance, Hamad Naeem et al. (2021) [19] propose a CNN-LSTM network for automated coronavirus detection, while Reagen L. Galvez et al. (2018) [18] focus on object detection using convolutional neural networks (CNNs). Additionally, Touseef Iqbal and Shaima Qureshi (2022) [17] discuss text generation models in deep learning. Research shows that combining CNNs with models like RCNN and Faster RCNN allows for effective visual information extraction and object detection, such as those by Zheng Zhang et al. (2019) [20] and Bin Liu et al. (2017) [21]. Finally, the production of explanatory text is assisted by using recurrent neural network (RNNs) and Long-Short Term Memory (LSTM), as demonstrated by PVC. Manogna et al. (2021) [22], the accuracy of image description is improved by integrating various strategies.

This study offers a simple captioning technique tailored for the quick-witted, high-stakes world of sports photography, as shown in “Smart Auto Image Captioning Using LSTM and Densenet Network" by Vijaykumar P et.al (2023) [23]. To develop a solid foundational understanding, the process begins with training a powerful deep learning network utilizing the Flickr8k dataset, which comprises of a range of random pictures, as observed in “Learning CNN-LSTM Architectures for Image Caption Generation" by Moses Soh (2016) [25]. A separate personal dataset has been chosen which contains images only related to sports, making sure it comprehends the smaller bits and other several complex sporting scenarios which is later fed to the model, as seen in “Image Caption Generator using CNN-LSTM " by Vishal Deshmukh et;al (2021) [24]. T Providing simply accurate, relevant caption along with representing sports or athletic scenes captured in picture is the main aim of this research project, as seen in “Image Caption Generator Using Densenet201 and Resnet50" by Vidhi Khubchandani (2024) [26]. The goal of this method is to increase the reliability and accuracy of automated image descriptions in the sports domain by focusing on the unique challenges associated with sports imaging, such as dynamic sceneries with a broad variety of actions, as seen in “Image Caption via Visual Attention Switch on DenseNet" by Yanlong Hao (2018) [27].

CNN and LSTM are the two algorithms applied to produce captions for pictures to understand text and images clearly. With this the strength of both the models is combined and its multiple layers and dense network help in extracting detailed visual features for which we have also used DenseNet201, as seen in “Comparison of Different CNN Model used as Encoders for Image Captioning" by Md Sahrial Alam et al; (2021) [28]. Following the passing of these attributes, an LSTM network processes the sequential data to produce captions that are accurate and consistent with the provided context, “Automatic image caption generation using deep learning" by Akash Verma et.al; (2023) [29]. As seen, the model generates high-quality picture descriptions appropriate for numerous applications like automatic image annotation and assistive technology by utilizing DenseNet201 for feature extraction and LSTM for sentence production, as observed in “A deep learning approach for classification of COVID and pneumonia using DenseNet-201" by Ankur Agarwal et.al; (2022) [30].

The remainder of the article is divided into the following sections: Section II of the Literature Review covers terminology, and Section III of that section includes pertinent material that elucidates the Experimental Setup. The approach is explained in depth in Section IV, and the findings and comments are presented in Section V. Section VI presents the evaluation that was done in Results and Discussion, and Section VII presents the conclusion.

## **Literature Review**

Automatic captioning of images has gained significant attention from academics and is considered an important task in computer vision. For algorithms to be able to describe images in a logical and contextually appropriate manner, advances in computer vision and natural language processing (NLP) would need to be merged.

**Early Approaches:** The majority of the initial methods for captioning images were template-based and comprised appending the characteristics of the items identified in the image to pre-established word patterns. Although these algorithms were able to produce basic captions, Initially, photo captioning techniques relied on templates, supplementing pre-made sentence patterns with information about the things identified in the image. While these techniques were able to produce basic captions, the sentences sounded forced and artificial most of the time. An extensive manual for evaluating models and information to write descriptions for photos taken by remote sensing. A detailed guideline for exploring models and data for remote sensing image caption generation was given by Lu and Wang (1998) [1].

**Deep Learning Advancements:** Deep learning has led to significant improvements in image captioning. Prioritizing the extraction of features from images, convolutional neural networks (CNNs) were employed, followed by recurrent neural networks (RNNs), specifically Long Short-Term Memory (LSTM) networks, for word generations. When these designs were integrated, the descriptions started to flow and become more organic. a thorough guide on model and data analysis for the generation of remote sensing picture captions, as discussed in a survey by Shuang Bai and Shah An (1989) [2].

**Encoder-Decoder Models:** The encoder-decoder framework is the usual method for caption creation for photographs. CNN is used for the initial analysis of the image, and RNN and LSTM are then used to provide the brief description. The LSTM condenses the summary that CNN delivers from the image into a sentence.This method is currently the industry standard for automatically creating image captions, as detailed by Alexander Toshev and Samy Bengio (1999) [3].

**Attention Mechanisms:** The algorithm learned where to look when creating each word in the caption because of these recent advancements in attention mechanisms. Because this approach provides a deeper comprehension of the data included in the image, it has enabled the provision of descriptions that are superior, as discussed by Ding, Qu, and Sangaiah (2019) [4].

**Datasets and Evaluation:** Several datasets have been important for training and testing picture captioning models, like Flickr8k, Flickr30k, and MS COCO. These files contain a large number of images, each with many user-written captions. The three most often utilized evaluation metrics are CIDEr, METEOR, and BLEU, as noted by Chen and Zitnik (2014) [5], are employed to evaluate these models' performance.

This paper also seeks to improve on these achievements through designing an image caption generator targeted at sport photos as this field faces unique challenges due to sports scenes’ dynamic nature and many angles.

**Research Gaps and Objectives**

The current study has highlighted several key knowledge gaps in the field of image captioning, stressing the gravity of need for substantial improvements particularly in scale, class and rotation ambiguity. However, despite rich insights contributed by the fusion of deep neural networks into template-based and retrieval-based approaches, limitations are evident in resultant sentences which necessitate further works. An issue that arises is that captions generated by these models do not correspond to the actual objects in picture fully; this indicates a requirement for more sophisticated methods that can correctly capture difficult scenes as well as describe them. These challenges need to be addressed if we are to improve the overall performance and precision of image captioning systems. In addition to identifying these research gaps, the study sets forth specific objectives aimed at addressing these shortcomings:

* Enhance feature localization within images to improve the accuracy of the image captioning process.
* Explore the integration of attention mechanisms to boost caption generation performance and coherence.
* Investigate the use of model ensembles to enhance the quality of generated captions.
* Contribute novel insights and methodologies to the field of image captioning.
* Strive towards producing more accurate and contextually relevant descriptions for diverse visual content.

**Experimental Setup**

To accomplish this research endeavor, an exhaustive attempt is made at constructing an image caption generator tailored for a wide dataset consisting of 50 sports pictures from different parts of the world. The aim is to give the model deeper learning competences so that it can generate captions that are more accurate and contextually suitable for such sports related images. The process done using python programming language takes place on Jupyter Notebook software. This setup gives you a hands-on coding experience with visuals, which is really important for making your model better over time.

This setup makes coding more fun with cool visuals, which helps make your model even better. TensorFlow and Keras are fundamental libraries to build and train the deep learning model, while Pillow helps us in image pre-processing tasks. Also, numpy handles all the numbers, and tqdm helps you see how your training is going. You can use these with JupyterLab, which makes trying things out easy and helps you develop your image captioning model faster.

The goal is to combine the two libraries into a single, comprehensive experiment to determine the best way to describe sports photos in the Corpus in a clear and precise manner. This should speed up developing captions for all kinds of sports photos automatically.

**Methodology**

Among the crucial phases in this study's methodology were data preparation, model training, and evaluation. We developed an efficient sports image caption generator using deep learning algorithms.

**Data Collection** : the dataset consists of sports images along with a text file containing captions for each image. Each image in the dataset has five different captions to ensure that each image has a wide range of descriptions that the model can train on.

1. **Loading and Cleaning Captions:** The captions were loaded, and all the punctuation marks and non-alphanumeric characters were removed. Then everything was converted to lowercase. After this the caption was saved for usage in the next step.
2. **Tokenization:** The third and most important step was giving each word in the revised captions a unique number value. This step holds significant importance as it prepares the textual information for input into the LSTM model.
3. **Padding:** In order to facilitate smooth batch processing while training the model, we needed to pad the tokenized sequences to ensure they all shared uniform length.

which aid in model training.

With each step, we dive- deeper into the world of sports imagery, unraveling its intricacies through text-based descriptions.

**Feature Extraction:** To extract features from the sports images, we used a pre-trained DenseNet201 model. This model was originally trained on the ImageNet dataset, but we adopted it to capture high-level features specific to sports images. The steps involved:

**1.Modifying the Model:** Adjustments were made to the DenseNet201 model by removing the final classification layers, thus preserving only the feature extraction layers.

**2.Generating Feature Vectors:** All images went through the updated DenseNet201 model to give out a 2048-dimensional feature vector, locking in crucial details about the image's content.

**3.Saving Features:** The extracted feature vectors were stored using a pickle file ensuring efficient retrieval

**Loading Dataset for Model Training:** The dataset file contained a list of image names used for training**.** Several functions were used to make the data preparation process easier and efficient:

* **File Parsing for Image Names:** A function was created to process a filename and generate a list of image names by parsing the text file’s contents.
* **Caption Dictionary Population:** Another function was created to populate a dictionary with captions corresponding to each image from the list of photos. To help the LSTM model identify the start and end of a caption, a unique identifier was added to each caption.
* **Feature Vector Retrieval:** Another function was implemented to retrieve the extracted feature vectors from the model, along with the dictionary of photo identifiers.

**Tokenizing the vocabulary:** In order to address the difficulty presented by machines' restricted understanding of sophisticated English terminology, the study constructs a more straightforward numerical representation of the model's input. The tokenizer function in the Keras library helps to assign a distinct index value to every word in the vocabulary. The final tokens are saved to a pickle file called "tokenizer.p" after being serialized. By vectorizing the text corpus and allocating indices to words, a tokenizer is used in this preparation stage to convert textual descriptions into a format appropriate for deep learning. In addition, the model architectural parameters are guided by the maximum length of descriptions. By means of these endeavors, the efficient employment of textual data in the deep learning framework is guaranteed, consequently augmenting the grasp and performance of the model.

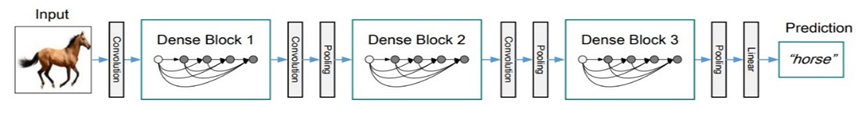


Figure 1: Overview of the CNN model DenseNet201 displaying steps of image recognition, leveraging dense connectivity for accurate predictions as detailed by Prof.S.Sankareswari et al., (2023) [6].

**Model Training:** A deep learning model that combines CNN for feature extraction and LSTM for sequence generation forms the basis of the picture caption generator. As illustrated in Figure 1, the model for creating captions for sports images is trained through a series of convolutional, pooling, and linear layers to provide the necessary results. The training process's several steps are as follows:

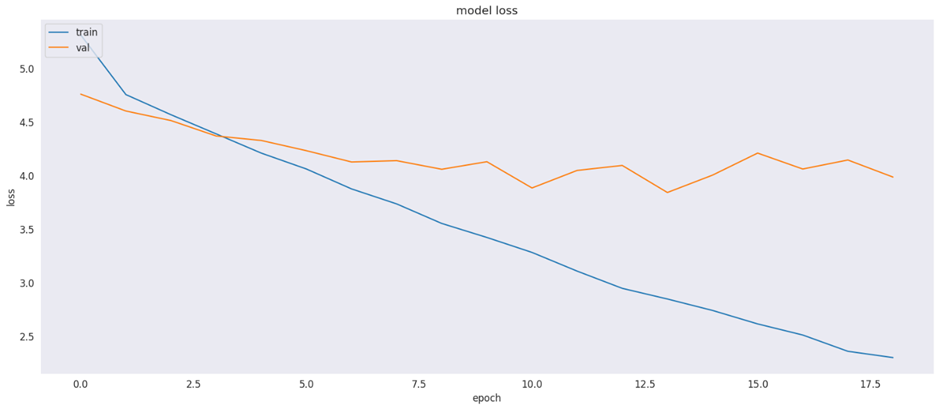
**1.Defining the Model Architecture:** Two encoders and one decoder are included in the model. The encoder used DenseNet201 to process and extract feature vectors from the images. The decoder used LSTM to construct captions based on these recovered features.

**2.Compiling the Model:** A model ideal for multi-class classification applications was constructed using the Adam optimizer and the categorical cross-entropy loss function.

**3.Training the Model:** Training the model required repeatedly exposing it to matched data. These were feature vectors—numerical representations—that were complemented by captions, which were written explanations. By making a few little adjustments, the model was able to match these pairs more accurately and with less error.

**Results and Discussion**

After training the sports photo caption model, there was a substantial difference in the model's performance on the train and test datasets. This discrepancy is shown in the graph below, which plots the loss against the total number of epochs.



*Figure 3: The above graph shows loss versus epochs during the training and testing of the Sports caption model.*

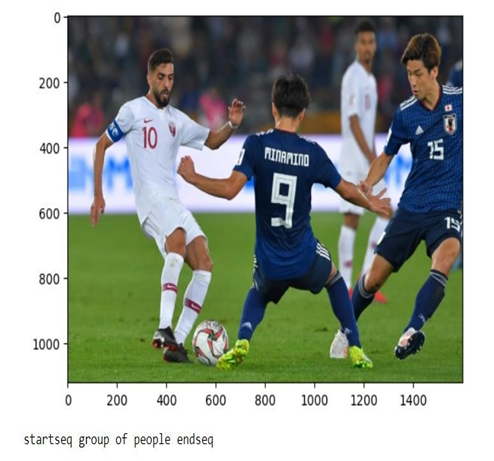
**Training Data Performance**

The model performed well on the training set, showing a steady decline in the evaluation measure or loss over the course of the training epochs. This implies that the model adjusts its parameters to reduce errors as it skillfully learns from the training set, refining its parameters to minimize errors.

**Test Data Performance**

However, the model's performance was not as good as the test data. Over different training epochs, the model continuously reduced the evaluation criterion which shows that the model performed well on the training fold. This shows how the model through learning from training data can adjust its parameters to reduce errors skillfully.

However different in terms of how well they performed, the model showed that it could produce captions with substance and emotion in them. Shown below is the caption given by model for sports picture. This example shows how well the model understands and describes complex sport scenes. Even with the performance variation across datasets, the model has the potential to make meaningful contributions to sports media and storytelling, descriptions.

*Figure 4: Generated Caption for Sports Image during the testing of Sports image caption generator model.*

**Conclusion and Future Work**

With success, this study has created a unique image caption generator designed especially for sports-related photos. The system utilizes a strong deep learning architecture that combines long short-term memory (LSTM) networks for sequence modeling and convolutional neural networks (CNNs) for feature extraction. Although the model performed well during training on a dataset of 50 sports photos, the study highlights an important finding: the importance of data bulk in improving caption accuracy. Business-grade models performed noticeably better after being trained on much bigger datasets—more than 100,000 photos. This result confirms the known connection between model accuracy and dataset size in deep learning applications.

To improve the accuracy, resilience, and general usefulness of sports picture caption generators, researchers have discovered a number of fascinating directions worth investigating in the future:

* The model can be improved to generate multiple captions from different points of view. This can help in providing more detailed info on the image from different perspectives.
* the models can be improved further to be able to describe the sports images better by understanding the context behind the image and the relationship between different objects in the image.
* The model can be improved further to provide captions in real time. This will be helpful during live sport event where captions need to be generated quickly to capture all the moments in a fast-paced game.

To sum up, this study shows a significant advancement in sports image captioning. It shows the importance of having a large dataset to improve the model's performance to produce a more accurate output. It sets a stage for more advanced methods to capture sports moments that can have real world applications.

**References**

[1] Kulkarni, G., Premraj, V., Dhar, S., Li, S., Choi, Y., Berg, A. C., & Berg, T. L. (2011). Baby talk: Understanding and generating simple image descriptions. In *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1601-1608).

[2] Vinyals, O., Toshev, A., Bengio, S., & Erhan, D. (2015). Show and tell: A neural image caption generator. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 3156-3164).

[3] 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).

[4] Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhutdinov, R., ... & Bengio, Y. (2015). Show, attend and tell: Neural image caption generation with visual attention. In *International Conference on Machine Learning* (pp. 2048-2057).

[5] Lin, T. Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., ... & Zitnick, C. L. (2014). Microsoft COCO: Common objects in context. In *European Conference on Computer Vision* (pp. 740-755). Springer, Cham.

[6] Sankareswari, S., Dongarkar, B. Z., Dongarkar, H., Sarang, S., & Valke, M. (2023). Image Caption Generator Using Deep Learning. FAMT, Ratnagiri, India.

[7] “Collecting image annotations using Amazon's Mechanical Turk” by Cyrus Rashtchian, Peter Young, Micah Hodosh, Julia Hockenmaier (2013).

[8] “Framing Image Description as a Ranking Task:Data, Models and Evaluation Metrics” by Micah Hodosh, Peter Young, Julia Hocken maier (2013).

[9] Aneja, J., Deshpande, A., & Schwing, A. G. (2018). Convolutional Image Captioning. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 5561-5570.

[10] Selivanov, A., Rogov, O. Y., Chesakov, D., Shelmanov, A., Fedulova, I., & Dylov, D. V. (2023). Medical Image Captioning via Generative Pretrained Transformers. [scientific reports].

[11] Allaouzi, I., Ben Ahmed, M., Benamrou, B., & Ouardouz, M. (2018). Automatic Caption Generation for Medical Images. [ACM digital library].

[12] Chohan, M., Khan, A., Mahar, M. S., Hassan, S., Ghafoor, A., & Khan, M. (2020). Image Captioning using Deep Learning: A Systematic Literature Review.

[13] Pan, X., Shi, J., Luo, P., Wang, X., & Tang, X. (2018). Spatial as Deep: Spatial CNN for Traffic Scene Understanding. AAAI-18/IAAI-18/EAAI-18 Proceedings, Volume One.

[14] Pan, J.-Y., Yang, H.-J., Duygulu, P., & Faloutsos, C. (2004). Automatic Image Captioning. IEEE International Conference.

[15] Chu, Y., Yue, X., Yu, L., Sergei, M., & Wang, Z. (2020). Automatic Image Captioning Based on ResNet50 and LSTM with Soft Attention. Hindawi.

[16] Zia, U., Riaz, M. M., & Ghafoor, A. (2022). Transforming Remote Sensing Images to Textual Descriptions. ScienceDirect, 108.

[17] Iqbal, T., & Qureshi, S. (2022). The Survey: Text Generation Models in Deep Learning. ScienceDirect, 34(6), Part A.

[18] Galvez, R. L., Bandala, A. A., Dadios, E. P., Vicerra, R. R. P., & Maningo, J. M. Z. (2018). Object Detection Using Convolutional Neural Networks. In TENCON 2018 - 2018 IEEE Region 10 Conference.

[19] Naeem, H., & Bin-Salem, A. A. (2021). A CNN-LSTM Network with Multi-Level Feature Extraction-Based Approach for Automated Detection of Coronavirus from CT scan and X-ray Images. ScienceDirect, 113(Part A).

[20] Zhang, Z., Wang, Y., Zhang, J., & Mu, X. (2019). Comparison of Multiple Feature Extractors on Faster RCNN for Breast Tumor Detection. In 8th ISNE.

[21] Liu, B., Zhao, W., & Sun, Q. (2017). Study of Object Detection Based on Faster R-CNN. In IEEE CAC, Volume 9.

[22] Buddana, H. V. K. S., Kaushik, S. S., Manogna, P. V. S., & Kumar, S. P. S. (2021). Word Level LSTM and Recurrent Neural Network for Automatic Text Generation. In IEEE ICCCI.

[23] Arumugam, M., Vijayakumar, P., Siddharrth, G. M., Rohini, J., Karthikeyan, A., & Bhat, C. Rohith. (2023). Smart Auto Image Captioning Using LSTM and Densenet Network. In 6th IEEE IC3I.

[24] Khant, P., Deshmukh, V., Kude, A., & Kiraula, P. (2021). Image Caption Generator using CNN-LSTM. IRJET, 08.

[25] Soh, M. Learning CNN-LSTM Architectures for Image Caption Generation. Stanford University.

[26] Khubchandani, V. (2024). Image Caption Generator Using Densenet201 and Resnet50. SSRN.

[27] Hao, Y., Xie, J., & Lin, Z. (2018). Image Caption via Visual Attention Switch on DenseNet. In IEEE IC-NIDC.

[28] Alam, M. S., Rahman, M. S., Hosen, M. I., Mubin, K. A., Hossen, S., & Mridha, M. F. (2021). Comparison of Different CNN Models Used as Encoders for Image Captioning. In IEEE ICDABI.

[29] Verma, A., Yadav, A. K., Kumar, M., & Yadav, D. (2023). Automatic Image Caption Generation Using Deep Learning. Springer Link, 83.

[30] Sanghvi, H. A., Patel, R. H., Agarwal, A., Gupta, S., Sawhney, V., & Pandya, A. S. (2022). A Deep Learning Approach for Classification of COVID and Pneumonia using DenseNet-201. WILEY.