Machines and Algorithms
*http://www.knovell.org/mna*

*Research Article,*

# Image caption generation using transfer learning using LSTM and DenseNet

**Abdul Jabbar**[1, *]

[1] Department of computer science, Riphah International University, I-14 Campus, Islamabad, 44000, Pakistan
[*]Corresponding Author: Abdul Jabbar. Email: abdul.jabbar1@riphah.edu.pk

Received: 30 Sept 2025; Revised: 29 Oct 2025; Accepted: 26 Nov 2025; Published: 21 Dec 2025
AID: 004-03-000058

**Abstract:** Image captioning consists of the description of images by identifying the main objects of an image, the features of the objects, and their associations. The effective system should also produce syntactically and semantically correct sentences. Deep learning methods can be effective in addressing the complications involved in this task. The article presents an advanced deep learning architecture of image captioning that enable the implication of three advanced technologies i.e., machine vision, machine translation and transfer learning. The state-of-the-art CNN architecture have been utilized to perform this task i.e., DenseNet201 model. DenseNet201 is a convolutional neural network (CNN) which converts the image data into a feature vector. After this CNN, a recurrent neural network (RNN) is exploited to encode the images using this vector. The coded text is then passed through another RNN, which is known as Long Short-Term Memory (LSTM) networks where the feature vector is decoded to produce a sequence of words which finally form the image descriptions. The Flickr8k dataset is used to test the effectiveness of the proposed model, and the performance of the model is measured with the help of the BLEU metric, which then gives a quantitative evaluation of the potential of the model.

**Keywords:** Image Captioning; DenseNet; Transfer Learning; Deep Learning;

## 1. Introduction

Image captioning describes the process of generating textual description for an image with the aid of advanced technologies like machine vision and natural language processing. This is done through a deep learning model to encode visual data and then decode it into a coherent paragraph or sentence. Image captioning is expected to produce a text sequence, which describes the input image [1]. These image captions are primarily based on deep neural networks and natural language processors to create a description of an image in terms of captions. Deep convolutional neural network (CNN) takes the input picture, codes it into a small number of features. Common CNN models used in this regard are VGG, VGG16, ResNet, DenseNet201 and Transformer-based models. An RNN (e.g. Long Short-Term Memory (LSTM) or Transformer-based) produces captions word-by-word. It gets conditioned on the image features as well as the words that it had previously generated [2][3].

Two primary methods of integrating image features in an image caption generator, early fusion and late fusion. In early fusion, image characteristics of various descriptors are concatenated in one vector. The image features are basically amalgamated at an early age usually before any language generation process has commenced. This method will encode both visual and linguistic information simultaneously in the initial stages, which can lead to memory-intensive models. The image features are added to the caption during the late fusion after the processing of the entire prefix of the caption. It delays the multimodal integration to a subsequent level. The neural language model, such as an RNN works using the caption prefix alone without considering the image features. In caption generation, the model uses visual information later, hence late fusion architectures are memory efficient [4]. The model that is the subject of discussion in this paper is the baseline captioning model [5], which has an encoder and a decoder. The encoder simultaneously obtains two image and text features. The text with encoder produces dense word representation of words in the embedding space. A CNN backbone is applied in the image encoder to produce high level image features. Decoder uses both image and text features to produce the image caption with LSTM.

The visual extraction of DenseNet and the modeling of LSTM are powerful, whereas the existing image captioning solutions do not promote the effective attention and multimodal fusion, which causes low visual-text correspondence. The study makes contributions such as (a) DenseNet transformer models and attention mechanisms are analyzed, (b) a hybrid learning framework to achieve better caption accuracy, and (c) increase image captioning robustly and avoid overfitting using publicly available datasets.

Rest of this research paper is ordered in the following way: The subsequent section will provide the literature review of the image captioning in depth. Section 3 illustrates the proposed model of the image captioning system. Section 4 will present the results, and the last one will be the conclusion of the paper, which is Section 5.

## 2. Literature Review

A lot of literature has been done on the image captioning and content generation of image captions [6-8]. The available image captioning schemes can be divided into two, namely, classical framework, comprising of template-based [9] and retrieval-based image captioning [10] and the second, encoder-decoder framework, comprising of pretrain models [11] and reinforcing learning models [12].

A new method used to solve vision-language problems is Oscar (Object-Semantics Aligned Pre-training). It takes question labels that are identified in pictures as the focal points to boost considerably the learning of classification. The disclosure that the brightest aspects in an image can be repeatedly recognized. Oscar is trained on 6.5M couples of text and images based open corpus and further refined on downstream tasks. This brings about new cutting-edge performance on six recognized tasks regarding vision-language understanding and generation. [13].

The study examines a new approach to the analysis of user-specific visual concepts in unstructured language, which is Personalized Vision & Language (PerVL). It presents benchmark datasets and suggests an architecture that can easily learn and use personalized visual concepts by extending pretrained models. The model is bigger, not only, but is more appropriate to VL processes. It has been trained with much larger pre-trained corpora, which contain publicly available annotated repositories of item identification data. This contrasts it with the top-down top-up models that are more popular. The method reveals that visual elements are important in visual language processes, unlike the past studies which ignored the original object detection model to improve the vision and language fusion model. Testing of the new object identification model incorporated the visual attributes in OSCAR, which is a transformer-based VL fusion model. This model was pre-trained and optimized with a better technique, OSCAR+, on several downstream VL model tasks. Recent progress in vision-language pre-training (VLP) has led to improvements in test results of the picture captioning challenge [14].

ClipCap is a new image captioning method that is useful in mining visual information of images. It applies a mapping network to produce a wide variety of context tokens, which helps to add to the greater accuracy of captioning and context understanding. The mapping network is either Multi-Layer Perceptron or Transformer which converts CLIP embedding to the language model space. This can be used to generate descriptions of images through training the language model using these context tokens. GPT-2 is trained to predict the following word in a caption, depending on first context tokens. They take less time because they can use a simpler architecture and ignore text encoder, only image encoder of CLIP is used [15].

Fine-grained image captioning with CLIP reward applies CLIP, a strong multimodal encoder, which has been trained on vast quantities of image-text data, to enhance the generation of captions. The given approach considers finer details and peculiarities of images, which leads to more detailed and correct captions. Rewarding with CLIP lets models create more detailed and unique descriptions without the use of reference captions to train. Finer-grained semantic rewards, including the caption reward and Semantic Segment Anything (SAM) reward, can potentially do much more to improve the fidelity of text prompts to generated images. This can therefore improve the visual quality and semantic similarity in text-to-image models. Image-based training is done in the CLIP-guided text GAN (CgT-GAN) approach, or CLIP-guided text GAN. This method improves the naturalness of captions and offers semantic guidance by training adversarially and using CLIP based rewards. Consequently, a tremendous performance change is realized in terms of different metrics of image captioning tasks [16].

Locality-Sensitive Transformer Network (LSTNet) is a new technique that aims at enhancing the local visual information processing by integrating Locality-Sensitive Attention and Fusion. This will allow the network to capture and use valuable visual information in a particular locality, which will result in the processing of data more effectively and accurately. The proposed system has three modules of designs such as LSA and LSF, LGC network introduces both short-range and long-range perception of the object features, JSAM extracts both global and local visual signals to form better sentences [17]. The proposed system consists of three components: 1) a detection or finding module, 2) a combination module, and a dedicated module for the generation of image captions.

1) Mask R-CNN [18], 2) YOLOv3 [19], and 3) RetinaNet [20] are all used in the detection module as an instance segmentation method. The combiner takes the outputs of two models i.e., YOLOv3 and Mask R-CNN initially. The output of the combination versus that of RetinaNet, the result of the combination is a result of three models on hierarchical levels of classes and superclasses. Image captioning is generally performed with the help of a strong deep learning framework which uses convolutional neural networks (CNNs) and recurrent neural networks (RNNs). This is a dynamic combination of features that allows the model to successfully analyse the complex visual characteristics of an image and provide a meaningful caption summarising what an image is about [21].

The proposed Bootstrapping Image-Text Alignment (BITA) method based on applying two-step vision-language pre-training approach and interactive fourier transformer is successful in aligning remote sensing images and text features, effectively outperforming other advanced methods on remote sensing image captioning tasks. The first step of BITA is based on contrastive learning of images and text to successfully match multiscale remote sensing image and text features. The second step of BITA is bridging the frozen image encoder and a large language model and uses prefix causal language modelling to steer the text generation procedure using visual features [22].

The authors present two new image captioning models, which consist of a knowledge retriever part, a differentiable encoder, and a kNN-enhanced language model. The given research presents the impressive effect of the explicit introduction of an external memory on the quality of captions, which becomes more pronounced with larger retrieval corpus. The authors give details on retrieval augmented captioning, enhancing the captioning of a large number of images [23].

ConvNeXt is being utilized as a feature extractor together with an LSTM block as well as a visual attention module. ConvNeXt exceeded the current standards of models with soft-attention and hard-attention systems by 43.04% and 39.04% respectively in BLEU-4 scores. It also did score higher in BLEU-4 score by 4.57% and 0.93% over vision transformers and data efficient image transformers respectively [24]. The authors introduce an ensemble image captioning model which involves a series of models with an attention mechanism in generating more precise captions. It boosts the visual-semantic congruence and concentrates on pertinent image regions when generating words by means of integrating different encoder-decoder designs [25].

## 3. Methodology

This part deals with the research methodology. The research is aimed at elaborating and accurate captions on images, through recognizing major ideas and concepts.

### *3.1. Feature Extraction*

The encoder derives visual features of an image. Encoders in general make use of convolutional neural networks (CNNs). It is usually followed as a visual recognition task model. All CNNs have four main layers. These layers consists upon a Convolutional layer for features extraction, the Pooling layer a down-sampling layer and a Flatten layer to transform dimension of extracted features and a fully connected layer for performing classification. A number of pre-trained CNN-based models are available to minimize the time required to train model.

DenseNet201 A CNN-based pretrained model, which includes a 201 layer deep convolutional neural network. It is also known to have complicated structure and connectivity patterns that are highly beneficial in propagating features in addition to gradient flow hence the model is best suited in different complex tasks. The size of DenseNet201 input is usually 224 x 224 pixels, the default size of the model. The input size of most deep learning models is usually 224x224 pixels (base paper).

### *3.2. Caption generation*

Caption generation model involves the use of RNN for the generation of intended sentences regarding input images. It is associated with the production of the feature extraction model. The simple RNN is ineffective when it is introduced to dealing with long chains of words. Nonetheless, it is possible to address this problem by integrating the Long Short-Term Memory (LSTM) network. LSTM structure uses memory cells. Typically, LSTM has been applied effectively in performing tasks like 1) sequence learning, 2) machine translation and 3) speech recognition which are also applicable to image captions.

LSTM was found to be useful in many applications like machine translation, sequence learning, speech recognition, and even image captioning. The memory cell and gates help to avoid gradient problems in LSTM. Figure 2 has three gates that are referred to as the input, output, and forget gates. A memory cell C controls these gates and has the duty of reading and writing. At time n, LSTM takes inputs in various sources. These are the present input (Xm), the past concealed state (Hm-1), and the preceding cell of memory state (Cm-1). The new values of the gate at time step n are calculated as follows:

$$I_m = \sigma(w_i. [H_{m-1}, X_n]) + b_I \quad (1)$$

$$\acute{C}_m = tanh(w_c. [H_{m-1}, X_n]) + b_c \quad (2)$$

$$F_m = \sigma(w_F. [H_{m-1}, X_m]) + b_F \quad (3)$$

$$O_m = \sigma(w_o. [H_{m-1}, X_n]) + b_o \quad (4)$$

$$C_m = F_m \times C_{m-1} + I_n \times \acute{C}_m \quad (5)$$

$$H_m = O_m \times tanh(C_m) \quad (6)$$

$I_m$, $F_m$, $O_m$ are the three gates i.e., including 1) input, 2) forget, 3) output gates and memory cell inputs respectively.  denotes values of weights, and denotes vectors of bias. The sigmoid activation function which is referred to as is as follows:

$$S_x = \frac{1}{1+exp(-X)} \quad (7)$$

The tanh activation function, denoted as $T_x$ , is calculated as:

$$T_x = \frac{exp(X)-exp(-X)}{exp(X)+exp(-X)} \quad (8)$$

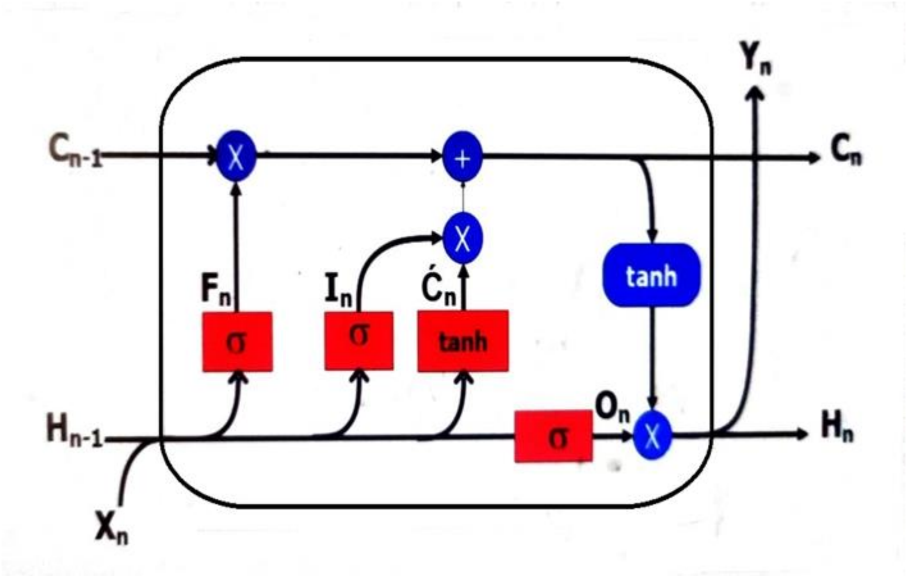


**Figure 1:** Block diagram of LSTM

The variable architecture is shown in Figure 2 and Table 1 depicts the parameters. The results of training are presented in Figure 3, and the loss of training and testing of the proposed architecture is shown in Figure 4.

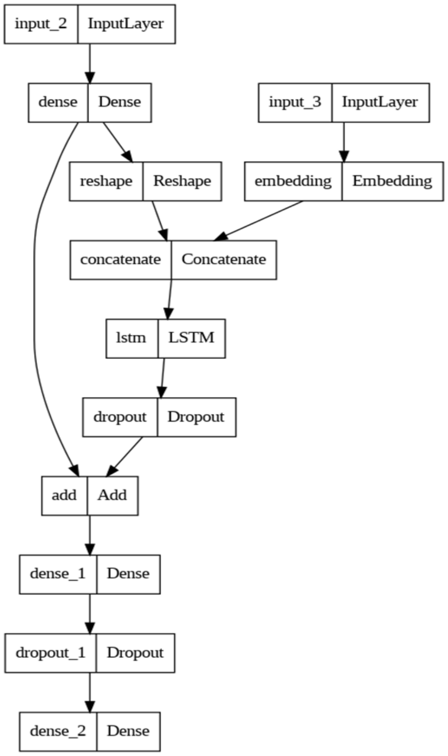
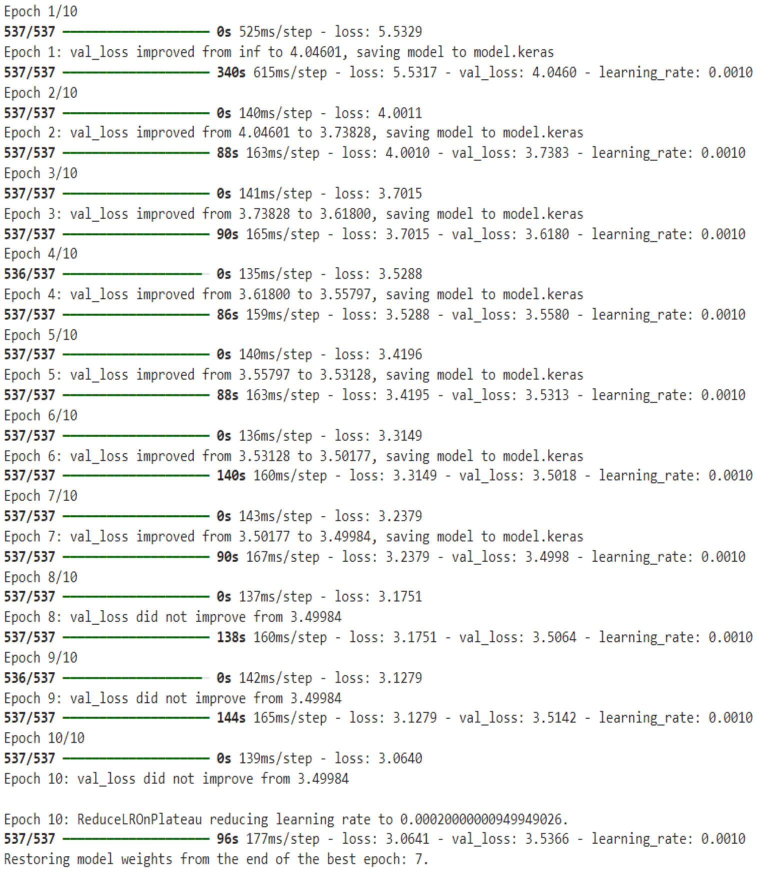


**Figure 2:** Proposed network for image captioning

**Table 1:** Parameter details of the proposed system

| Component | Layer Name | Type | Output Shape |
|---|---|---|---|
| **Caption Input** | input_layer_3 | Input | (None, 34) |
| **Image Input** | input_layer_2 | Input | (None, 1920) |
| **Word Embedding** | embedding_1 | Embedding | (None, 34, 256) |
| **Image Feature Encoder** | ImageFeature | Dense | (None, 256) |
| **Caption Encoder** | CaptionFeature | LSTM | (None, 256) |
| **Feature Fusion** | add_1 | Add | (None, 256) |
| **Fully Connected Layer** | dense_2 | Dense | (None, 256) |
| **Output Layer** | dense_3 | Dense | (None, 8,485) |



```
Epoch 1/10
537/537 ━━━━━━━━━━━━━━━━━━━━ 0s 525ms/step - loss: 5.5329
Epoch 1: val_loss improved from inf to 4.04601, saving model to model.keras
537/537 ━━━━━━━━━━━━━━━━━━━━ 340s 615ms/step - loss: 5.5317 - val_loss: 4.0460 - learning_rate: 0.0010
Epoch 2/10
537/537 ━━━━━━━━━━━━━━━━━━━━ 0s 140ms/step - loss: 4.0011
Epoch 2: val_loss improved from 4.04601 to 3.73828, saving model to model.keras
537/537 ━━━━━━━━━━━━━━━━━━━━ 88s 163ms/step - loss: 4.0010 - val_loss: 3.7383 - learning_rate: 0.0010
Epoch 3/10
537/537 ━━━━━━━━━━━━━━━━━━━━ 0s 141ms/step - loss: 3.7015
Epoch 3: val_loss improved from 3.73828 to 3.61800, saving model to model.keras
537/537 ━━━━━━━━━━━━━━━━━━━━ 90s 165ms/step - loss: 3.7015 - val_loss: 3.6180 - learning_rate: 0.0010
Epoch 4/10
536/537 ━━━━━━━━━━━━━━━━━━━╺ 0s 135ms/step - loss: 3.5288
Epoch 4: val_loss improved from 3.61800 to 3.55797, saving model to model.keras
537/537 ━━━━━━━━━━━━━━━━━━━━ 86s 159ms/step - loss: 3.5288 - val_loss: 3.5580 - learning_rate: 0.0010
Epoch 5/10
537/537 ━━━━━━━━━━━━━━━━━━━━ 0s 140ms/step - loss: 3.4196
Epoch 5: val_loss improved from 3.55797 to 3.53128, saving model to model.keras
537/537 ━━━━━━━━━━━━━━━━━━━━ 88s 163ms/step - loss: 3.4195 - val_loss: 3.5313 - learning_rate: 0.0010
Epoch 6/10
537/537 ━━━━━━━━━━━━━━━━━━━━ 0s 136ms/step - loss: 3.3149
Epoch 6: val_loss improved from 3.53128 to 3.50177, saving model to model.keras
537/537 ━━━━━━━━━━━━━━━━━━━━ 140s 160ms/step - loss: 3.3149 - val_loss: 3.5018 - learning_rate: 0.0010
Epoch 7/10
537/537 ━━━━━━━━━━━━━━━━━━━━ 0s 143ms/step - loss: 3.2379
Epoch 7: val_loss improved from 3.50177 to 3.49984, saving model to model.keras
537/537 ━━━━━━━━━━━━━━━━━━━━ 90s 167ms/step - loss: 3.2379 - val_loss: 3.4998 - learning_rate: 0.0010
Epoch 8/10
537/537 ━━━━━━━━━━━━━━━━━━━━ 0s 137ms/step - loss: 3.1751
Epoch 8: val_loss did not improve from 3.49984
537/537 ━━━━━━━━━━━━━━━━━━━━ 138s 160ms/step - loss: 3.1751 - val_loss: 3.5064 - learning_rate: 0.0010
Epoch 9/10
536/537 ━━━━━━━━━━━━━━━━━━━╺ 0s 142ms/step - loss: 3.1279
Epoch 9: val_loss did not improve from 3.49984
537/537 ━━━━━━━━━━━━━━━━━━━━ 144s 165ms/step - loss: 3.1279 - val_loss: 3.5142 - learning_rate: 0.0010
Epoch 10/10
537/537 ━━━━━━━━━━━━━━━━━━━━ 0s 139ms/step - loss: 3.0640
Epoch 10: val_loss did not improve from 3.49984

Epoch 10: ReduceLROnPlateau reducing learning rate to 0.00020000000949949026.
537/537 ━━━━━━━━━━━━━━━━━━━━ 96s 177ms/step - loss: 3.0641 - val_loss: 3.5366 - learning_rate: 0.0010
Restoring model weights from the end of the best epoch: 7.
```

**Figure 3:** Sample of Training Results of the Proposed System

## 4. Results

Different datasets are important in image captioning methods training, testing and evaluation. Such datasets are highly heterogeneous in terms of images, captions/image, caption format and captions/image. Some of the most commonly used data are Flickr8k [26], Flickr30k [27] and the MS COCO Dataset [28].

One of the most known data sets is Flickr8k [29], which basically incorporates 8K images. These images are collected on Flickr. For the training of model, we have dedicated 6K images, while for testing and validation of them 1K images are dedicated for each task. Interestingly, in this dataset, there are five reference captions that are annotated with each image as depicted in Figure 6. Many techniques of image captioning have used this dataset to perform experiments [30].

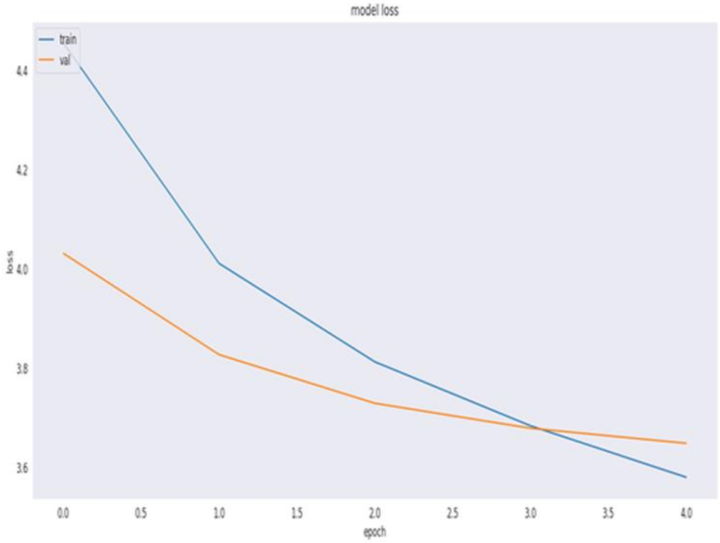
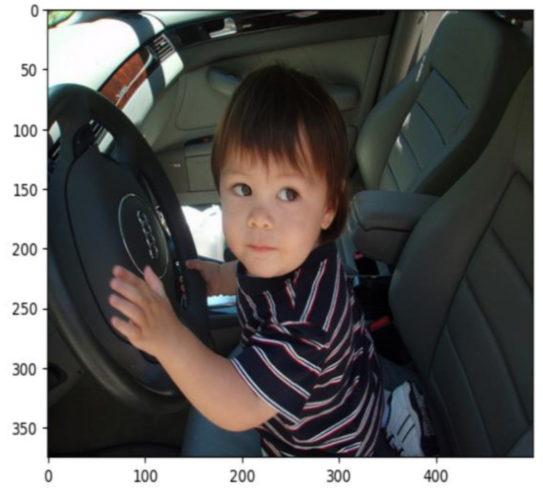


**Figure 4: Demonstration of training and testing loss of the proposed system**

DenseNet-with-LSTM is able to reuse more features and propagate gradient, rendering more vivid visual representations than CNNs do in [5]. Although both of them use LSTM to generate language, DenseNet provides smaller image features, enhancing relevance and accuracy of captions. Table 2 mentions the blue score outcome of the suggested system.



```
['a little boy at the steering wheel of a vehicle ',
 'a small baby sitting in a car behind the steering wheel',
 'a young boy pretends to drive the car  with his hand on the horn ',
 'little boy sits in car s driver s seat  grabbing the steering wheel ',
 'this small child is sitting behind the steering wheel of a car ']
```

**Figure 5:** Sample output of the proposed system

**Table 2: Comparison of results of the system**

| References | Caption generation approach | BLEU Score |
|---|---|---|
| **Proposed** | Hybrid | **0.71750** |
| **[5]** | DenseNet201 | 0.70750 |

## 5. Conclusion

As the findings discussed in the paper explain, the revised structure of the Densenet201 model does not only record a very high BLEU score but also indicates that the produced captions are of a very high quality. Such high BLEU score indicates the effectiveness of the model in terms of the ability to reflect the complex visual information presented in the images. It, therefore, highlights the ability of this model to successfully coded the visual information to produce textual descriptions which are understandable and contextually viable. In the total, the results support the possibilities of this modified architecture to improve image captioning work.

Implementing the state-of-the-art attention-based models to extract image features will enable the system to perform its tasks much better and provide real-time automatic captioning of video streams.

**Funding Statement:** No external funds received.

**Conflicts of Interest:** NIL.

**Data Availability:** Flickr8k dataset is publicly available.

## References

[1] Sen, A. (2023). Captioning Image Using Deep Learning Approach. International Journal for Research in Applied Science and Engineering Technology.

[2] Muhammad Shah, F., Humaira, M., Jim, M. A. R. K., Saha Ami, A., & Paul, S. (2022). Bornon: Bengali image captioning with transformer-based deep learning approach. SN Computer Science, 3, 1-16.

[3] Castro, Roberto, Israel Pineda, Wansu Lim, and Manuel Eugenio Morocho-Cayamcela. (2022). Deep learning approaches based on transformer architectures for image captioning tasks. IEEE Access, 10, 33679-33694.

[4] Tanti, Marc, Albert Gatt, and Kenneth P. Camilleri. (2018). Where to put the image in an image caption generator. Natural Language Engineering, 24(3), 467-489.

[5] Vinyals, Oriol, Alexander Toshev, Samy Bengio, and Dumitru Erhan. (2015). Show and tell: A neural image caption generator. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 3156-3164).

[6] Salgotra, G., Abrol, P., & Selwal, A. (2025). A survey on automatic image captioning approaches: Contemporary trends and future perspectives. Archives of Computational Methods in Engineering, 32(3), 1459-1497.

[7] Thobhani, A., Zou, B., Kui, X., Abdussalam, A., Asim, M., Shah, S., & Elaffendi, M. (2025). A Survey on Enhancing Image Captioning with Advanced Strategies and Techniques. Computer Modeling in Engineering & Sciences (CMES), 142(3).

[8] Jamil, A., Mahmood, K., Villar, M. G., Prola, T., Diez, I. D. L. T., Samad, M. A., & Ashraf, I. (2024). Deep learning approaches for image captioning: Opportunities, challenges and future potential. IEEE Access.

[9] Rahman, I. U., Wang, Z., Liu, W., Ye, B., Zakarya, M., & Liu, X. (2020). An N-state Markovian jumping particle swarm optimization algorithm. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 51(11), 6626-6638.

[10] Liu, S., Xian, Y., Li, H., & Yu, Z. (2017). Text detection in natural scene images using morphological component analysis and Laplacian dictionary. IEEE/CAA Journal of Automatica Sinica, 7(1), 214-222.

[11] Verma, A., Yadav, A. K., Kumar, M., & Yadav, D. (2024). Automatic image caption generation using deep learning. Multimedia Tools and Applications, 83(2), 5309-5325.

[12] Ren, Z., Wang, X., Zhang, N., Lv, X., & Li, L. J. (2017). Deep reinforcement learning-based image captioning with embedding reward. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 290-298).

[13] Li, X., Yin, X., Li, C., Zhang, P., Hu, X., Zhang, L., ... & Gao, J. (2020). Oscar: Object-semantics aligned pre-training for vision-language tasks. In Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16 (pp. 121-137). Springer International Publishing.

[14] Zhang, P., Li, X., Hu, X., Yang, J., Zhang, L., Wang, L., ... & Gao, J. (2021). Vinvl: Revisiting visual representations in vision-language models. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 5579-5588).

[15] Mokady, R., Hertz, A., & Bermano, A. H. (2021). Clipcap: Clip prefix for image captioning. arXiv preprint arXiv:2111.09734.

[16] Cho, J., Yoon, S., Kale, A., Dernoncourt, F., Bui, T., & Bansal, M. (2022). Fine-grained Image Captioning with CLIP Reward. Findings of the Association for Computational Linguistics: NAACL 2022.

[17] Ma, Yiwei, Jiayi Ji, Xiaoshuai Sun, Yiyi Zhou, and Rongrong Ji. (2023). Towards local visual modeling for image captioning. Pattern Recognition, 138, 109420.

[18] He, Kaiming, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. (2017). Mask r-cnn. In Proceedings of the IEEE international conference on computer vision (pp. 2961-2969).

[19] Redmon, Joseph, and Ali Farhadi. (2018). Yolov3: An incremental improvement. arXiv preprint arXiv:1804.02767.

[20] Li, Yixing, and Fengbo Ren. (2019). Light-weight retinanet for object detection. arXiv preprint arXiv:1905.10011.

[21] Rinaldi, Antonio M., Cristiano Russo, and Cristian Tommasino. (2023). Automatic image captioning combining natural language processing and deep neural networks. Results in Engineering, 18, 101107.

[22] Yang, Cong, Zuchao Li, and Lefei Zhang. (2024). Bootstrapping interactive image-text alignment for remote sensing image captioning. IEEE Transactions on Geoscience and Remote Sensing.

[23] Sarto, Sara, Marcella Cornia, Lorenzo Baraldi, Alessandro Nicolosi, and Rita Cucchiara. (2024). Towards Retrieval-Augmented Architectures for Image Captioning. ACM Transactions on Multimedia Computing, Communications and Applications.

[24] Ramos, Leo, Edmundo Casas, Cristian Romero, Francklin Rivas-Echeverría, and Manuel Eugenio Morocho-Cayamcela. (2024). A study of convnext architectures for enhanced image captioning. IEEE Access.

[25] Al Badarneh, I., Hammo, B. H., & Al-Kadi, O. (2025). An ensemble model with attention based mechanism for image captioning. Computers and Electrical Engineering, 123, 110077.

[26] Micah Hodosh, Peter Young, and Julia Hockenmaier. 2013. Framing image description as a ranking task: Data, models and evaluation metrics. Journal of Artifcial Intelligence Research 47 (2013), 853–899.

[27] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015.Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In Proceedings of the IEEE international conference on computer vision. 2641–2649.

[28] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr DollÃąr, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In European conference on computer vision. Springer, 740–755

[29] Micah Hodosh, Peter Young, and Julia Hockenmaier. 2013. Framing image description as a ranking task: Data, models and evaluation metrics. Journal of Artifcial Intelligence Research 47 (2013), 853–899

[30] Hossain, M. Z., Sohel, F., Shiratuddin, M. F., & Laga, H. (2019). A comprehensive survey of deep learning for image captioning. ACM Computing Surveys (CsUR), 51(6), 1-36.