Pertanika J. Sci. & Technol. 34 (S1): 89 - 110 (2026)

PERTANIKA JOURNALS

SCIENCE & TECHNOLOGY

Journal homepage: http://www.pertanika.upm.edu.my/

Enhanced Image Captioning Using Sinusoidal Spatial Transform CNN and Fine-tuned BERT

Bhargavi Polepalli*, Praveen Kumar Sekharamantry, and Konda Srinivasa Rao

Department of Computer Science and Engineering, GITAM School of Technology, GITAM (Deemed to be University), Gandhi Nagar, Rushikonda, 530045 Visakhapatnam, Andhra Pradesh, India

ABSTRACT

Image caption generation is a significant area of study in artificial intelligence and computer vision, focusing on training systems to generate accurate textual descriptions of images. This paper presents an advanced framework for image captioning using a dataset sourced from Kaggle. The process begins with pre-processing techniques, including the Flexiframe Filter, which dynamically adjusts the window size based on local variance to reduce noise in both text and images. Bright-Contrast augmentation is then applied to enhance input images, enriching the dataset and improving feature extraction. The encoder module utilises a Sinusoidal Spatial Transform Convolutional Neural Network (SST-CNN) integrated with a Visual Geometry Group (VGG16) model for feature extraction and encoding. For the decoder, stochastic k-sampling is combined with a Dense Neural Network and a Gated Recurrent Unit (GRU) to generate initial captions. To refine these captions, a fine-tuned Bidirectional Encoder Representations from Transformers (BERT) model is employed for enhanced coherence and accuracy. The model's performance is evaluated using standard metrics, achieving BLEU-1, METEOR, ROUGE-L, and CIDER scores of 1, 0.99, 0.99, and 1, respectively. The proposed approach demonstrates significant improvements in generating descriptive and accurate image captions, making it a robust solution for applications requiring semantic understanding of visual data.

Keywords: Bright and contrast augmentation, fine-tuned BERT, Flexiframe Filter, Sinusoidal Spatial Transform Convolutional Neural Network (SST-CNN), Stochastic k-sampling DenseGRU Neural Network

ARTICLE INFO

Article history:
Received: 07 March 2026
Accepted: 29 April 2026
Published: 24 July 2026

DOI: https://doi.org/10.47836/pjst.34.S1.05

E-mail addresses:
bpolepal@gitam.in (Bhargavi Polepalli)
psekharm@gitam.edu (Praveen Kumar Sekharmantry)
skonda@gitam.edu (Konda Srinivasa Rao)
* Corresponding author

INTRODUCTION

Image caption generation can surely be termed as an interesting and complex problem in the new and constantly

e-ISSN: 2231-8526

developing fields like artificial intelligence and computer vision. Its goal is to create techniques and algorithms through which these textual descriptions of the images may be created (Varma et al., 2026). Stated that this skill relates to the translation of language to vision, and it has a profound effect on the availability, delivery, and interaction of information to a computer.

The Significance of Image Caption Generation

The ability to come up with accurate and culturally appropriate descriptions of images is important in diverse industries. It is essential to point out that the accessibility of image captions is decisive because people who are blind or have low vision can get an understanding of images via textual descriptions (Beddiar et al., 2023). Functions such as image captioning methods, for example, can be better at describing complex scenes, items, and activities and also enhance the use of social media sites, news websites, and educational sites.

Automated captioning systems can also help to enhance the indexing and categorisation of a substantial number of visual contents when used in the management of digital media (Mahalakshmi & Fatima, 2022). This is particularly useful for several industries such as e-commerce that may benefit from making product searches even more effective, while customers may be more engaged by the meaningful captions (Afyouni et al., 2021). The suitable captioning of the shown content means that the material recommendation, suggestion, and personalisation functions in the field of media and entertainment become more effective.

Applications and Implications

The process of caption writing can be useful in a wide array of areas. It can be beneficial to make images more readable, especially for people with some sort of vision issues, since automatic image captions help enhance this aspect. Successful and interesting captions can improve the overall experience as well as product descriptions for e-commerce stores (Degadwala et al., 2021). Caption generation can help in the process of personalisation and the creation of content in social media and digital marketing. In addition, the development of an automatic captioning system must lead to the improvement of the processes of language acquisition, as well as help with assistive technologies in education (Hossain et al., 2021).

Even though image caption generation has made significant progress, there remain several challenges and opportunities, the exposure of which is the following: generating captions for images depicting complex environments and understanding conceptual opinions and cultural and contextual relevance (Al Duhayyim et al., 2022). It is also important to anticipate and solve the ethical concerns for using image captioning systems that include bias and fairness in this technology's development and deployment (Xian et al., 2022).

This involves applying the SST-CNN method to generate reasonable as well as informative captions and computer vision approaches to extract substantial objects from images.

In Section 2, we present a literature review including datasets, findings, and image captioning methods. In Section 3, we present our approaches for producing a variety of captions. Sections 4 present the outcomes of the experiment. Section 5 is the conclusion

LITERATURE REVIEW

In this section, evaluation image captioning methods, dataset, and their evaluation metrics are shown in Table 1.

Table 1
Literature review description

Objective	Dataset	Findings
Puscasiu et al. (2020) conducted extensive field research and attempted to create an application for deep learning-based automated image description.	80000 annotated images from Microsoft Common Objects in Context (MSCOCO)	To help visually impaired persons make sense of images, the technology had to be further connected with a text-to-speech engine.
In the field of computer vision research, addressed by Castro et al. (2022). Visual attention is a cutting-edge method for image captioning.	MSCOCO 2014 dataset	For the BLEU-4 and Top-5 accuracy, MobileNetV3 supports values of 19.50 and 72.928, respectively, indicating that it provided competitive output quality at the expense of reduced computing performance.
Devi et al. (2020) numerous models, including the cutting-edge encoder-decoder (Sequential CNN-Recurrent Neural Network (RNN) systems that had shown to be effective in producing outcomes, had been put forth to address the issue.	They make use of the Flickr8k and 30k databases.	Comparing their performance to other models, they obtained state-of-the-art results on both datasets.
Wang et al. (2022), presented a novel approach to image captioning using recognised elements and relationships through high-order interaction learning within the encoder-decoder architecture.	MSCOCO dataset	Numerous tests demonstrated that the suggested approach could outperform cutting-edge techniques on the MSCOCO database in a competitive manner.
The efficacy of the Approach was demonstrated in the research by (Al-Malla et al., 2022). Presenting an attention-based Encoder-Decoder image captioning model that makes use of two feature extraction techniques: an object detection module, an image classification CNN (Xception), and You Only Look Once version 4 (YOLOv4).	MS COCO and Flickr30k	The CIDER score was increased by 15.04% with their new feature extraction technique.

Table 1 (*continued*)

Objective	Dataset	Findings
Shen et al. (2020) proposed a two-stage multitask learning system depending on reinforcement learning (RL) and a variational autoencoder. For remote sensing image captioning, the two-stage Multi-task Learning Model (VRTMM) relied on reinforcement learning and a variational autoencoder.	NWPURESISC45 dataset	The experiment's outcome showed that their model works well for remotely sensed image captioning and produced unique, cutting-edge results.
The Task-Adaptive Attention module for image captioning presented by Yan et al. (2022) could help with the deceptive issue and teach implicit non-visual	MSCOCO captioning Dataset	Comprehensive tests on the MSCOCO captioning database showed that performance improvement was possible when their Task-Adaptive Attention module was plugged into a vanilla Transformer-based image captioning model.

RESEARCH METHODOLOGY

Image captions are generated using the following methodology: gathering data from the Kaggle source, preprocessing images with text and Flexiframe Filter and Bright-Contrast augmentation, feature extraction with SST-CNN and VGG16, producing initial captions with stochastic k-sampling, dense neural networks, and GRU and enhancing captions with fine-tuned BERT. The evaluation employs the BLEU, METEOR, CIDER, and ROUGE-L metrics to evaluate caption integrity and efficiency. Figure 1 shows the overview of the methodology.

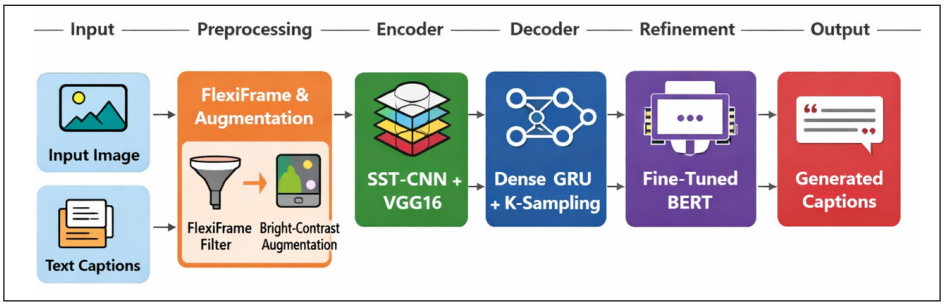

Figure 1. Overview of methodology

Data Collection

8,000 images with five distinct captions that clearly describe the important objects and events are included in these unique benchmark collections for sentence-based image description and examination. The images were picked from six separate Flickr categories, and they

were carefully chosen to show a range of settings and scenarios rather than any famous persons or places. We used 4000 distinct images from the Flickr8k dataset gathered from the Kaggle source https://www.kaggle.com/datasets/adityajn105/flickr8k?select=Images. Figure 2 shows the sample images.

Figure 2. Sample images

Data Preprocessing

This improves the model's capacity to capture context-specific details and produce more precise and expressive captions by dynamically shifting the frame of attention to focus on important image regions. Preprocessing text entails cleaning, tokenising, shrinking the vocabulary, padding sequences, and adding sequence tokens. Grayscale conversion, Gaussian blur, edge detection, thresholding, contour detection, and mask application are some of the image preprocessing techniques used with the Flexiframe Filter to improve feature extraction.

Text Preprocessing

Text cleaning eliminates punctuation, single-character words, numeric values, and extraneous characters, making the characters cleaner so that important words can be recognised. Add Sequence Tokens at the start and finish of each caption, including special tokens $('startseq' \, and \, 'endseq')$ aids in identifying the initial and finish of sequences in systems requiring these markers. Text data is prepared for neural network input by tokenisation, which turns words into numerical values and turns captions into sequences of integers where each word is mapped to a distinct integer. Calculate the Size of The Vocabulary using text analysis, determine the vocabulary size required to cover a given percentage of the data (e.g., 95% coverage), and decide how many terms should be added to the vocabulary that are unique. Pad Sequences use zeros to buffer shorter sequences so that they are all the same length, creating uniformly sized sequences for feeding neural networks.

Flexiframe Filter

Grayscale Conversion reduces the image to a single channel and turns it to grayscale to facilitate processing. Gaussian Blur enhances edge recognition by reducing noise and smoothing the grayscale image. Edge Detection: Sobel Operator finds edges in the x- and y-axes to draw attention to the image's key elements. Thresholding and Contour Detection generate a binary image and use contour identification to separate important features. Mask Application determines which sections inside the identified contours should be retained by applying the binary mask to the original image.

Each image has been filtered to highlight and isolate important details. Along with a final message stating that the processing is finished, the console shows the names of every processed image. Figure 3 displays the resultant preprocessed image.

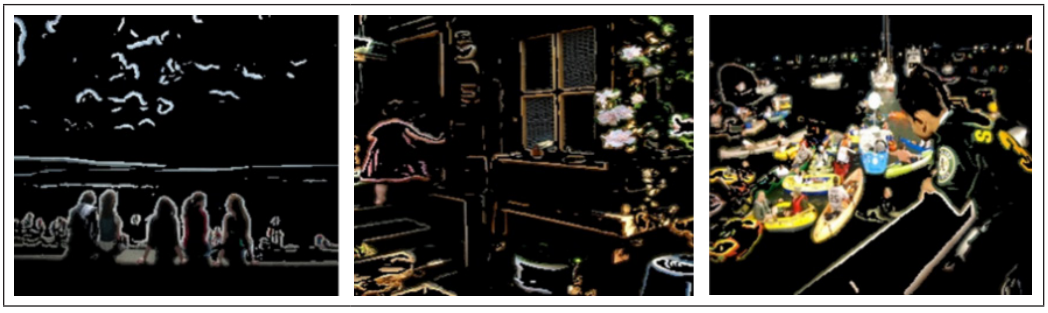

Figure 3. Output of preprocessed images

Augmentation Using Bright-Contrast

Strengthening the models' resistance to visual changes, this approach aids in the improvement of image captioning efficiency. Brightness Adjustment uses a brightness factor of 1.2 to increase brightness by 20%. This scales pixel values to brighten the image. Contrast Adjustment applies a contrast factor of 0.8 to reduce contrast by 20%. Modifying the mean intensity and scaling the pixel values lowers the contrast. Send back the updated picture with the made changes.

Images that have been modified using particular augmentation settings make up the output:

- A 20% improvement in brightness was seen (brightness factor of 1.2).
- A 20% drop in contrast was seen (contrast factor: 0.8). Figure 4 shows the output of augmented images.

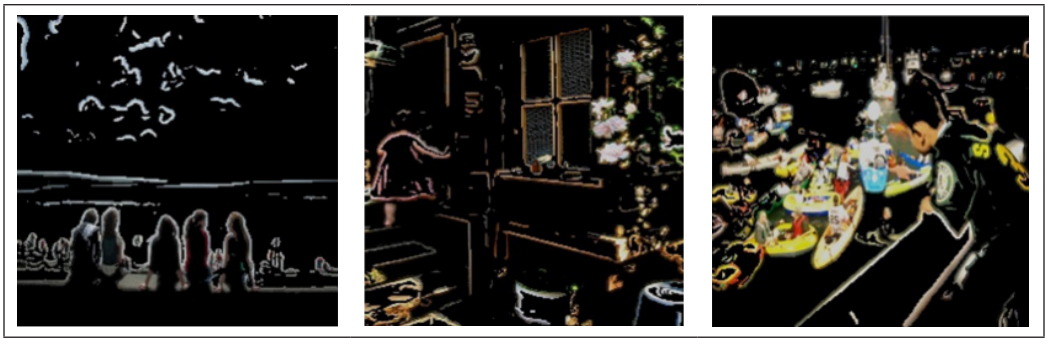

Figure 4. Output of augmented images

Image Feature Extraction Using Pre-trained Model VGG16

These features help to provide meaningful and contextually appropriate captions for images by acting as input for image captioning systems. To recognise patterns and visualise, data must initially be converted into numerical values, a process known as feature extraction. It is a necessary phase that aids the model in determining greater precision. The most crucial element in choosing a suitable feature extraction and processing method is the abundance of varied variables present in these enormous collections. Processing these variables requires a substantial amount of computational resources. To efficiently decrease data size, the process of feature extraction involves choosing and integrating parameters to create features that extract the greatest possible information from images.

The VGG16 feature extraction approach is capable of extracting large amounts of data with better performance. This method is widely employed for extracting features from images and has a substantial impact. The VGG16 model is applied to a collection of image data. To extract features, the proposed models' designs use a variety of layers across their architecture. The complete framework was built using convolutional, batch normalisation, pooling, and fully connected layers, as well as ReLU and SoftMax functions. The VGG16 is a compact design that includes a pooling layer, convolution, and a fully connected layer, all of which are present in the AlexNet model. There are an entire 16 layers that expand with the number of layers. This network keeps the input image size at 224 by 224 pixels, with a 3 by 3 filter. The final step of this network includes an activation function effective at giving class possibilities to the output layer collected image features by the VGG16 pre-trained model displayed in Figure 5.

```
Extracted Features:

Features for 1048710776_bb5b0a5c7c.jpg:
[0.         0.         0.65566266 ... 0.29483652 0.60420024 0.        ]

Features for 1042020065_fb3d3ba5ba.jpg:
[0.25006348 0.         0.         ... 0.         0.         1.6714716 ]

Features for 1000268201_693b08cb0e.jpg:
[1.6208034 2.9046288 0.        ... 0.        0.        2.1787658]
```

Figure 5. Extracted image features by VGG16 pre-trained model

Encoder Using SST-CNN

This technique improves image captioning by providing richer spatial context for generating accurate and descriptive captions.

The CNN is a sophisticated model; its classification and prediction accuracy will decrease with a particular alteration or interference of the target image. The data sets are all thought to be mixed-written digital images based on the concept of model optimisation. An SST-CNN was developed as a new network structure based on the CNN model. By identifying and aligning anomalous and conspicuous regions in the image, the SST can rectify and align. Below are the particular steps involved in network design.

Stage 1

Establish the network for placement. As an input, the feature map (FM)V is utilised. It is possible to employ a convolutional or fully connected structure. Outputting the spatial transformation component θ and recording $\theta = floc(V)$ is the purpose.

Stage 2

Convert the incoming FM using the spatial transformation component θ that the positioning network outputs. Using the linear transformation as an instance, it may map to a place on the input map with the component (w_j^t, z_j^t) and output a specific position on (w_j^s, z_j^s). The computation Equation 1 is as follows.

$$\begin{pmatrix} w_j^t \\ z_j^t \end{pmatrix} = \theta \begin{pmatrix} w_j^s \\ z_j^s \\ 1 \end{pmatrix} = \begin{bmatrix} \theta_{11} & \theta_{12} & \theta_{13} \\ \theta_{21} & \theta_{21} & \theta_{21} \end{bmatrix} \begin{pmatrix} w_j^s \\ z_j^s \\ 1 \end{pmatrix} \quad [1]$$

Stage 3

Use bilinear interpolation to complete the spaces as per the completing rules specified in the preceding period. For completing, the pixel value equivalent to the coordinate point in V is derived using the coordinate point of U, negating the requirement for matrix operations.

Note that there is no direct performance of the completion. Once some processing has been done, the additional pixel values surrounding the completion should be taken into consideration if the calculated coordinates are decimals. Here is the complete preparation in Equation 2.

$$U_j = \sum_n \sum_m V_{mn} * l\left(w_j^t - n; \emptyset_w\right) * l\left(z_j^t - m; \emptyset_z\right) \quad [2]$$

The sampling kernels $\emptyset_w$ and $\emptyset_z$ are shown above. The parameter L determines the type of image interpolation. Here, the input FM V_{mn} is V position. $U\ (m, n)$, Location is $\left(w_j^s, z_j^s\right)$ and U_j is the output FM. The following Equation 3 can be used to apply bilinear interpolation.

$$U_j = \sum_m \sum_n V_{mn} * \max\left(0,1 - \left|w_j^t - n\right|\right) * \quad \max\left(0,1 - \left|z_j^t - m\right|\right) \quad [3]$$

Encoder Output

- The encoder in an autoencoder transfers the input data to a latent space, usually one with fewer dimensions. Put another way, the encoder produces an encoding or compressed description of the input. The result highlights the most significant aspects of the supplied data while removing the less crucial information.
- The encoder output is typically significantly more compact and efficient than the input since its dimensionality is substantially smaller.

Decoder Using Stochastic k-Sampling DenseGRU Neural Net

The model decodes the latent features into the original image dimensions using a unique technique that combines DenseGRU layers with stochastic k-sampling.

Stochastic k-Sampling

To improve the diversity and the potential of the generated captions in the environment of image caption decoding, stochastic k-sampling can be used. A model is trained to produce descriptive sentences for images in image captioning. This usually requires an encoder-decoder architecture, in which the decoder (usually an RNN or Transformer) creates a string of words to make a caption, and the encoder (usually a CNN) drives features from the image.

Top-k Sampling

The top k probable terms in the distribution are chosen by the model, as opposed to the word with the highest probability. The number of top words that are taken into consideration is

determined by the hyperparameter k. The top 5 probable terms are taken into account by the model if $k = 5$ Using these top k candidates and their probability to determine the following word, stochastic sampling adds randomisation.

Stochastic Selection

The top k words are chosen at random, but the probability of selecting each word is proportionate to the probability that the model predicts. For instance, there is a 50% probability of selecting the first word, a 30% chance of selecting the second, and a 20% chance of selecting the third if the probabilities of the top three words are 0.5, 0.3, and 0.2.

Diversity in Captions

By adding unpredictability to the decoding procedure, stochastic k-sampling facilitates the creation of several unique captions for a single image. Every time the model decodes a caption using a different set of stochastic options, it might generate a distinct and legitimate caption. This variety is especially useful for tasks where an image can have several reasonable interpretations.

Trade-off

The ratio of correctness to diversity is balanced. The captions could become less accurate or logical if k is set too high. The captions may get monotonous and less varied if k is set too low. The particular application and intended results determine which value of k to use.

Dense GRU Neuro Net

A GRU layer processes the sampled features, efficiently capturing complex dependencies and patterns in the latent space. Subsequently, the GRU layer's output is fed into a variety of dense layers during the dense expansion stage. These layers gradually enlarge the data to encompass the entire image, utilising regularisation strategies and non-linear transformations to improve the model's efficiency and capacity for generalisation.

GRU

To the GRU network e_h, the feature input tensor is delivered. Depending on the level of hyperparameter effectiveness, the GRU could have multiple layers (ML) or perhaps one layer. $W[s] \in \mathbb{R}^{m_c \times g}$ Represents the input for a given time step s, and the following computations are made, as shown in Equations 4 to 7.

$$Q[s] = \sigma\big(W[s]X_{wq} + G[s-1]X_{gq} + a_q\big) \quad [4]$$

$$Y[s] = \sigma\big(W[s]X_{wy} + G[s-1]X_{gy} + a_y\big) \quad [5]$$

$$\tilde{G}[s] = \tanh\big(W[s]X_{yg} + (Q[s] \odot G[s-1])X_{gg} + a_g\big) \quad [6]$$

$$G[s] = Y[s] \odot G[s-1] + \big(1 - Y[s] \odot \tilde{G}[s]\big) \quad [7]$$

Where the hidden state (HS)of the previous time stage is represented by $G[s-1] \in \mathbb{R}^{m_c \times g}$, the reset gate is represented by $Q[s] \in \mathbb{R}^{m_c \times g}$, the update gate is represented by $Y[s] \in \mathbb{R}^{m_c \times g}$, the weight variables are $X_{wq} X_{wy} \in \mathbb{R}^{m_c \times g}$ and $X_{gq} X_{gy} \in \mathbb{R}^{g \times g}$, the biases are represented by $a_q a_y a_g \in \mathbb{R}^{1 \times g}$, and the number of features in the layer is represented by g $(g = m_e)$ for the first layer). The new state is $G[s]$, and the candidate concealed state is $\tilde{G}[s]$~ stands for the Hadamard product $\odot$ and σ for the sigmoid function. The input of a specific layer in an ML GRU network is the HS $G[s]$ of the layer before it. There isn't a dropout utilised among the layers.

Dense GRU Neuro Net

Concurrently, the feature tensor is restructured into a vector $W_{c1} \in \mathbb{R}^{m_c \times (m_e * 24)}$, which is supplied to a solitary dense layer network e_{c1}. The activation function that is employed is the Rectified Linear Unit (ReLU). The dense layer can be expressed as follows, assuming that $(Z_{c1})^l$ is the output vector and $(W_{c1})^l$ is the input vector for the l^{th} developing instance in Equations 8 and 9.

$$U_j^l = ReLU\big(\sum x_{ji} (w_{c1})_i^l - ag_j\big) \quad [8]$$

$$(Z_{c1})_o^l = \sum U_j^l x_{oj} - ap_o \quad [9]$$

Where U_j^l the weight component from the hidden layer neuron j to the input layer neuron i is the variance of the output neuron ag_j is the partiality of the hidden neuron $(\mathrm{HN})j$, and x_{oj} is the weight component from $\mathrm{HN}j$ to output o. The outputs Z_{c1} from the dense layer and Z_h from the final GRU layer are connected.

A network of attention e_b receives the connected layer $W_b(Z_h, Z_{c1})$. There's a utilisation of generous assistance. Below is an explanation of the attention mechanism.

Initially, a collection of internal states called $g_1, g_2, \ldots, g_n$ are created by encoding the input connected sequence. A feedforward network is trained to identify important

states by assigning greater ratings to states that merit attention, and conversely. This process yields alignment ratings for every encoded state. The softmax function is used to calculate attention weights $\propto_1, \propto_2, \ldots\ldots \propto_n$ for the ratings. Take note that the attention weight vector, with $\propto \in [0,1]$ and $\sum \alpha = 1$, provides a probabilistic interpretation. The setting vector is then calculated in Equation 10.

$$D = \propto_1 * g_1 + \propto_2 * g_2 + \cdots .. + \propto_n * g_n \quad [10]$$

The output from the earlier time stage is connected with D. For every time step, the procedure is performed. Using a second dense layer network e_{c2}, the desired drift sequence Δ,which once more employs ReLU as the activation function, is assigned to the attention layer output Z_b. Thus, the DGNN model $e_{DGNN}: W \rightarrow \Delta$ can be summed up as follows in Equations 11 to 14.

$$GRU\ e_h: \quad W \rightarrow Z_h \quad [11]$$

$$Densel\ e_{c1}: W_{c1} \rightarrow Z_{c1} \quad [12]$$

$$Attention\ e_b: W_b(Z_h, Z_{c1}) \rightarrow Z_b \quad [13]$$

$$Dense2 \quad e_{c2}: W_{c2}(Z_b) \rightarrow \Delta \quad [14]$$

The following summarises the formulas for each activation function that was employed in the model shown in Equations 15 to 18.

$$\sigma(w) = \frac{1}{(1+e^{-w})} \quad [15]$$

$$\tanh(w) = \frac{(e^{w} - e^{-w})}{e^{w} + e^{-w}} \quad [16]$$

$$ReLU(w) = \max(w, 0) \quad [17]$$

$$softmax_j(\vec{w}) = \frac{e^{w_j}}{\sum_{i=1}^{n} e^{w_i}}\ for\ j = 1, \ldots\ldots, n \quad [18]$$

Generated Caption [Fine-tuned BERT]

The BERT model is fine-tuned to handle the task of generating captions by learning from pairs of images and their corresponding textual descriptions. This process fine-tuned BERT's ability to produce meaningful and coherent sentences based on the visual content represented by the images.

Ensuring optimal performance on numerous natural language processing tasks, the multi-headed attention model BERT was among the first to employ the encoder portion of the initial transformer architecture. K Encoder layers coupled in series, each with G attention heads utilised in parallel, make up this model.

Since the attention elements of BERT $_{\text{BASE}}$ $(K = 12, and\ G = 12)$ are the portion of the model that contains the topological data that are of interest in analysing, concentrate on this. A matrix W of size $m \times c$ (where $c = 512$) containing vector illustrations of the m tokens in the input phrase serves as the attention head's input. Its output is a novel representation W^{out} of size $m \times c$ with $c' < c$ $(c' = 64)$ using the following Equation 19.

$$W^{out} = X^{attn}.(W.X^{U}) with\ X^{attn} = softmax(\frac{(W.X^{R}).(W.X^{L})^{S}}{\sqrt{c}} \qquad [19]$$

Here, the token embedding is projected onto a lower-dimensional vector space via trainable vectors $X^R,\ XL, X^U \in \mathbb{R}^{c \times c'}$, and the softmax is applied over the second dimension called the attention matrix, the matrix X^{attn} is a $m \times m$-dimensional grid.

When calculating the novel description of the token at location j, its entries x_{ji}^{attn} can be understood as the attention provided to the token at location j. They are non-negative, and each token's attention scores add up to one; therefore, for every $\sum_{i=1}^{m} x_{ji}^{attn} = 1 \quad j = 1, \ldots., m.$

Display Output (Images with Captions)

- The last stage involves manually entering images from the dataset as inputs from the user, and our model creates descriptions for each of these images.
- The final product will consist of image-caption pairs, where each image has a detailed caption produced by the optimised BERT model. This enables users to evaluate the model's efficacy in producing pertinent and cohesive descriptions based on the input images both visually and textually.
- Additionally, we create captions for arbitrary images found online. Figure 6 shows the output of images with captions.

Figure 6. Output of an image with a caption

RESULTS AND DISCUSSION

The experimental environment and setup, as well as the effectiveness of the suggested approach displayed in Table 2, are covered in this section and compared with the existing approaches, which are deep convolutional neural networks and recurrent neural networks with attention (CNN-RNN-Att) (Barati et al., 2023), Residual Network with 50 layers (ResNet-50) (Khaing et al., n.d.).Table 3 shows the overall performance.

Table 2
Experimental setup

Experimental Setup	Details
Model	SST-CNN with dense GRU neural net and fine-tuned BERT
Task	Image Caption Generation
Dataset	Collected from Kaggle sources
Hardware	Laptop running Windows 11
RAM	16 GB
Software Environment	Python 3.10.1
Evaluation Metrics	BLEU-1, METEOR, CIDER, ROUGE-L

Table 3
Overall result comparison

Methods	BLEU-1	METEOR	ROUGE-L	CIDER
DCNN-RNN-Att [16]	0.786	0.291	0.579	0.713
ResNet50 [17]	0.576	0.207	0.213	0.454
SST-CNN [Proposed]	1	0.99	0.99	1

BLEU-1

To determine how well the generated text matches descriptions made by humans, it computes the ratio of matching terms between the created caption and reference captions. Figure 7 shows the comparative evaluation of BLEU-1.

While the existing methods DCNN-RNN-Att and ResNet-50 achieved 0.786 and 0.576, respectively, our proposed SST-CNN methodology achieved (1). The findings demonstrate that our suggested approach outperforms existing methods substantially.

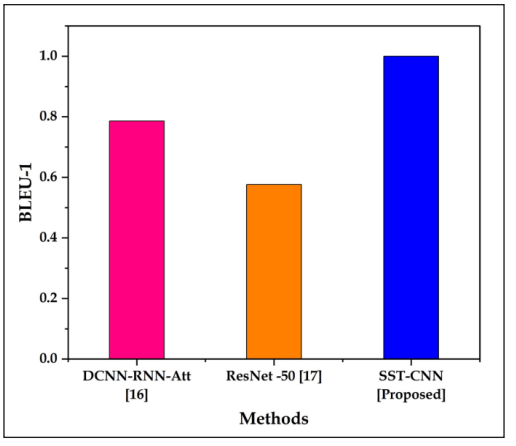

Figure 7. Output of BLEU-1

METEOR

METEOR scores are calculated based on word and phrase matches in the produced and reference captions to provide a more accurate and relevant evaluation. Figure 8 shows the comparison assessment of METEOR.

While the existing methods DCNN-RNN-Att and ResNet-50 achieved 0.291 and 0.207, respectively, our proposed SST-CNN methodology achieved 0.99. The results show that our proposed strategy performs significantly better than current approaches.

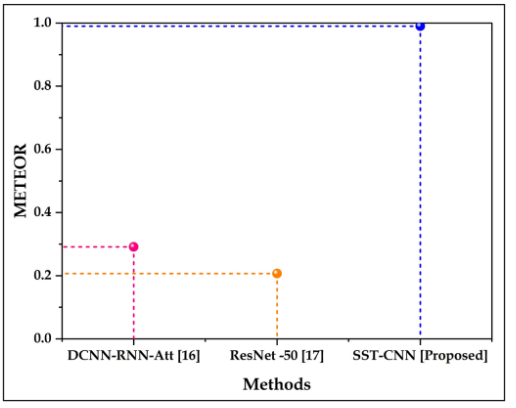

Figure 8. Output of METEOR

CIDER

It assesses the consistency and importance of words and phrases, emphasising details and environment to determine caption descriptive quality. Figure 9 shows the comparison of CIDER.

While the existing methods DCNN-RNN-Att and ResNet-50 achieved 0.713 and 0.454, respectively, our proposed SST-CNN methodology achieved (1). The outcomes demonstrate that, compared to existing methods, our suggested solution works noticeably better.

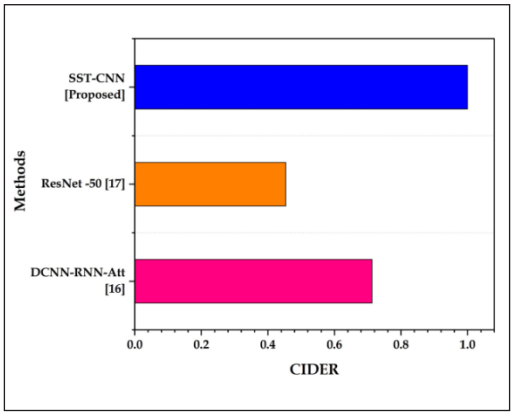

Figure 9. Output of CIDER

ROUGE-L

It represents the quality and relevancy of captions by taking into account both content overlap and sequence order, resulting in meaningful and coherent descriptions. The ROUGE-L score is displayed in Figure 10.

While the existing methods DCNN-RNN-Att and ResNet-50 achieved 0.579 and 0.213, respectively, our proposed SST-CNN methodology achieved 0.99 for Image caption generation. The outcomes demonstrate that, compared to existing methods, our suggested solution works substantially well.

Figure 11 shows how similar metrics BLEU, METEOR, ROUGE-L, and CIDEr have performed in various models. The suggested SST-CNN model has a higher score than that of DCNN-RNN-Att and ResNet50, which illustrates its usefulness in caption generation tasks. The drastic rise in performance is indicative of enhanced semantic comprehension and contextual correspondence. These results are consistent

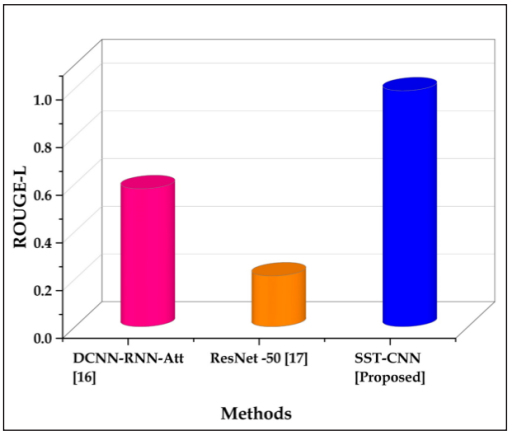

Figure 10. Output of ROUGE-L

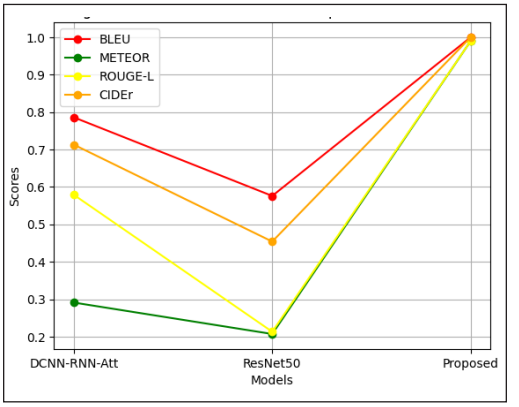

Figure 11. Comparison of performance metrics through the use of a line chart

with the previous research that highlights the importance of hybrid deep learning models (Alkhaldi et al., 2025).

The bar chart in Figure 12 gives a comparative evaluation of metrics grouping of the models, with the proposed approach prevailing. All metrics show a significant change, especially METEOR and CIDEr, suggesting the accuracy of language and contextual understanding. The visualisation is useful in highlighting the performance difference between the conventional CNN-RNN models and the hybrid architecture. The same enhancement has been experienced when using attention-based captioning systems (Xu et al., 2026).

This scatter plot as shown in Figure 13 will be used to visualise the distribution of the metric scores of each of the models, allowing one to see the dispersion of performance. The points on the proposed model are concentrated at higher values, which means stable and better results on all the metrics. Conversely, the variation and poor consistency of performance are observed in the baseline models. These patterns of distribution are typical of deep learning capturing systems (Dsouza et al., 2026).

The box plot as shown in Figure 14 is the statistical data of performance measures such as median, quartiles and variability. The model proposed has greater median values with less variation, which shows consistent performance. The spreads in the baseline models are broader, indicating instability in the quality of prediction. This statistical image lends credence to the strength of the SST-CNN-based structure (Dsouza et al., 2026).

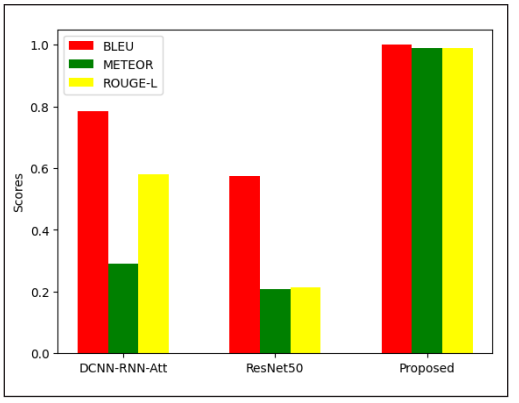

Figure 12. Relative analysis of metrics with bar chart

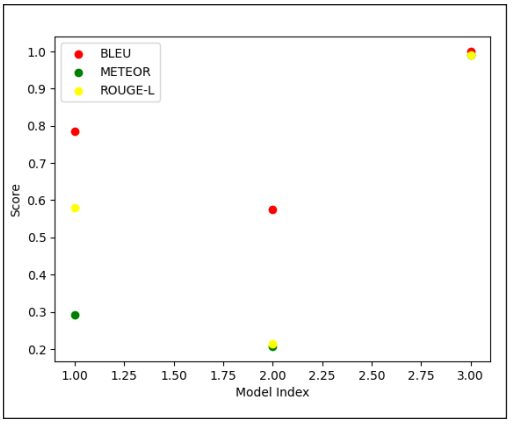

Figure 13. Scatter plot of model performance metrics

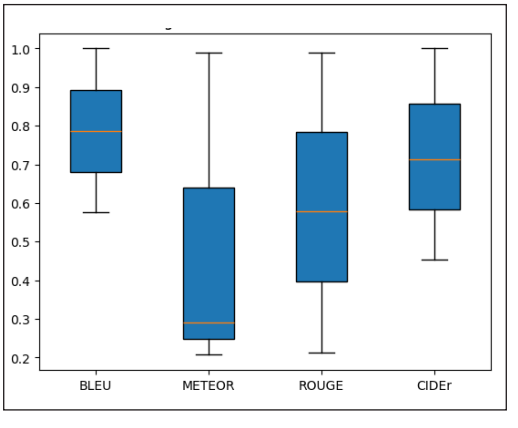

Figure 14. Evaluation metrics distribution with box and whisker plot

The heatmap, as shown in Figure 15, is an easy visualisation of the intensities of metrics across models in the form of colour gradients. The more intense areas are associated with improved performance, which is a clear indication of the superiority of the proposed method. The equal values of all metrics show balanced performance. A heatmap is commonly employed to assess visual tasks on deep learning models (Dsouza et al., 2026).

The radar chart as shown in Figure 16 is a multi-dimensional comparison of evaluation metrics which demonstrate the overall performance profile of each model. The suggested model is almost a complete polygon, which means that all metrics are of equal scores. On the contrary, baseline models exhibit irregular shapes, indicating uneven performance. The graph is a good example of the balanced transformation that the suggested architecture can bring (Dsouza et al., 2026).

This value depicts variability and confidence levels of the BLEU scores of the models as shown in Figure 17. The model proposed has a low error margin, implying that it is very reliable and stable in its predictions. Baseline models exhibit bigger variations, indicating inconsistency in generated captions. The error bar analysis is essential in verifying model strength in experimental studies(Dsouza et al., 2026).

The bubble chart as shown in Figure 18 represents the BLEU scores and the relative

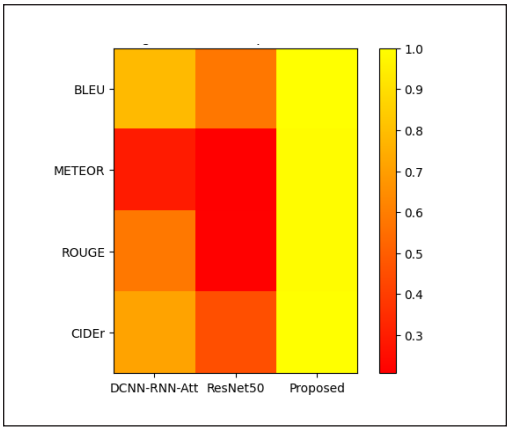

Figure 15. Heatmap view of metric scores

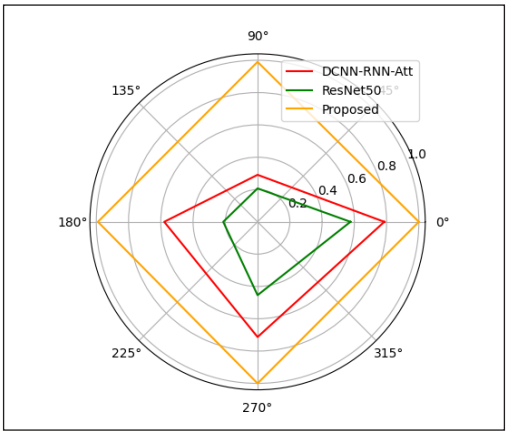

Figure 16. Multi-metric multi-measure radar chart

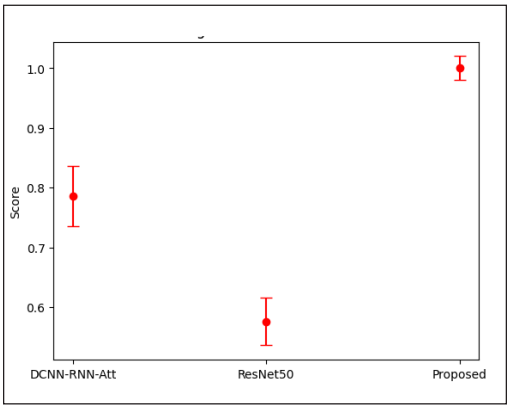

Figure 17. BLEU scores analysis as an error bar

importance of the score in terms of different sizes of the bubble. The largest bubble with the highest score represents the proposed model and highlights that it is the best. This visualisation is a combination of magnitude and value, providing a total comparison. These representations come in handy in emphasising predominant models in comparative studies (Dsouza et al., 2026).

The Gantt chart, as shown in Figure 19, is used to describe the consecutive steps of the suggested methodology, such as preprocessing, feature extraction, training, and evaluation. It gives a definite chronological insight into the workflow and computational pipeline. This formal representation increases the understandability of the system design. In complicated deep learning models, workflow visualisation is critical (Dsouza et al., 2026).

The flow of computational resources in the Sankey diagram is illustrated by the various steps in the model, including feature extraction as shown in Figure 20, GRU processing, and BERT refinement. Flow thickness is the contribution to the computation made by each stage. This diagram makes the resource-consuming aspects of the architecture obvious. Such analysis is important for optimising deep learning pipelines (Dsouza et al., 2026).

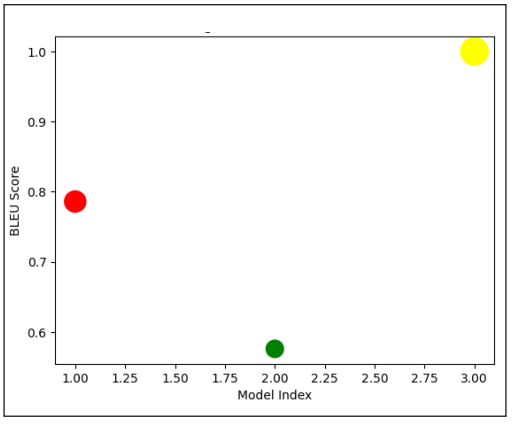

Figure 18. Bubble chart of distribution of BLEU score

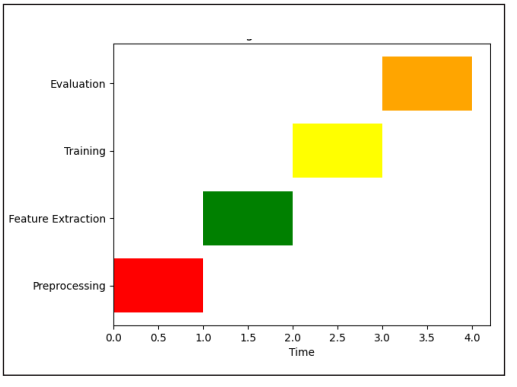

Figure 19. Workflow representation Gantt chart

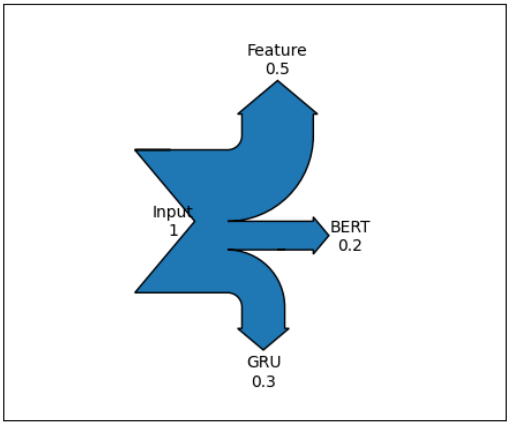

Figure 20. Computational flow Sankey diagram

CONCLUSION

In the present work, we have presented a way to generate captions for images using the SST-CNN system. We acquired a dataset from Kaggle, which we processed post-acquisition through both Flexiframe Filter text and image consistency as well as the Bright-Contrast method to greatly boost images for augmentation. The SST-CNN model is very effective in feature extraction and encoding, with the help of the VGG16 model. Decoder using stochastic k-sampling with a Dense Neural Network and GRU, and we used caption generation to fine-tune the BERT method. Evaluation of the model was done using the BLEU-1 (1), METEOR (0.99), ROUGE-L (0.99), and CIDER (1) to ensure the model's capacity to generate accurate and quality captions. It may encounter obstacles that could impair efficiency and scalability in practical applications, such as increased processing complexity, noise sensitivity, challenges managing a variety of image settings, limited generalisability across different datasets, and possible inefficiencies while processing complicated or high-resolution images; these are some of the drawbacks that may be encountered. Subsequent endeavours will focus on refining computing efficiency, strengthening noise resilience, and expanding flexibility in various image scenarios. Furthermore, using multimodal inputs and sophisticated attention mechanisms could improve the relevance and accuracy of captions even more. Further research is needed in the area of enhancing the efficiency and scalability of computations, so that image captioning in resource-constrained environments can be realised in real time. Moreover, further integration of multimodal architectures and attention-based transformers can be used to improve the accuracy of captions and contextual retrieval. Generalising the model to various datasets and multilingual captioning cases will enhance generalisation and practical usefulness.

ACKNOWLEDGEMENT

Author Contributions: All authors contributed equally, and all authors read and approved the final version of the paper.

REFERENCES

Afyouni, I., Azhar, I., & Elnagar, A. (2021). AraCap: A hybrid deep learning architecture for Arabic image captioning. *Procedia Computer Science, 189*, 382-389. https://doi.org/10.1016/j.procs.2021.05.108

Al Duhayyim, M., Alazwari, S., Mengash, H. A., Marzouk, R., Alzahrani, J. S., Mahgoub, H., Althukair, F., & Salama, A. S. (2022). Metaheuristics optimisation with deep learning-enabled automated image captioning system. *Applied Sciences, 12*(15), Article 7724. https://doi.org/10.3390/app12157724

Alkhaldi, T. M., Asiri, M. M., Alzahrani, F., & Sharif, M. M. (2025). Fusion of deep transfer learning models with Gannet optimisation algorithm for an advanced image captioning system for visual disabilities. *Scientific Reports, 15*, Article 40446. https://doi.org/10.1038/s41598-025-24171-9

Al-Malla, M. A., Jafar, A., & Ghneim, N. (2022). Image captioning model using attention and object features to mimic human image understanding. *Journal of Big Data, 9*, Article 20. https://doi.org/10.1186/s40537-022-00571-w

Barati, A., Farsi, H., & Mohamadzadeh, S. (2023). Integration of the latent variable knowledge into deep image captioning with Bayesian modelling. *IET Image Processing, 17*(7), 2256-2271. https://doi.org/10.1049/ipr2.12790

Beddiar, R., & Oussalah, M. (2023). Explainability in medical image captioning. In A. Holzinger, R. Goebel, R. Fong, T. Moon, K.-R. Müller, & W. Samek (Eds.), *xxAI: Beyond explainable AI: International Workshop, Held in Conjunction with ICML 2020, July 18, 2020, Vienna, Austria, Revised and Extended Papers* (pp. 239-261). Academic Press. https://doi.org/10.1016/B978-0-32-396098-4.00018-1

Castro, R., Pineda, I., Lim, W., & Morocho-Cayamcela, M. E. (2022). Deep learning approaches based on transformer architectures for image captioning tasks. IEEE Access, 10, 33679-33694. https://doi.org/10.1109/ACCESS.2022.3161428

Degadwala, S., Vyas, D., Biswas, H., Chakraborty, U., & Saha, S. (2021). Image captioning using Inception V3 transfer learning model. In 2021 6th International Conference on Communication and Electronics Systems (ICCES) (pp. 1103-1108). IEEE. https://doi.org/10.1109/ICCES51350.2021.9489111

Devi, P. R., Thrivikraman, V., Kashyap, D., & Shylaja, S. S. (2020). Image captioning using reinforcement learning with BLUDEr optimisation. Pattern Recognition and Image Analysis, 30(4), 607-613. https://doi.org/10.1134/S1054661820040094

DSouza, R., & Sen, S. (2026). A SwinBERT framework with temporal windowed cross attention for video captioning for visual language intelligence. Discover Applied Sciences.

Hossain, M. Z., Sohel, F., Shiratuddin, M. F., Laga, H., & Bennamoun, M. (2021). Text-to-image synthesis for improved image captioning. IEEE Access, 9, 64918-64928. https://doi.org/10.1109/ACCESS.2021.3075579

Khaing, P. P., San, M., & Aung, M. M. (n.d.). Analysis on residual network models for image description generation. Journal of Research and Applications, 1(1), 208-213.

Mahalakshmi, P., & Fatima, N. S. (2022). Summarisation of text and image captioning in information retrieval using deep learning techniques. *IEEE Access, 10*, 18289-18297. https://doi.org/10.1109/ACCESS.2022.3150414

Puscasiu, A., Fanca, A., Gota, D. I., & Valean, H. (2020). Automated image captioning. In *2020 IEEE International Conference on Automation, Quality and Testing, Robotics (AQTR)* (pp. 1-6). IEEE. https://doi.org/10.1109/AQTR49680.2020.9129930

Shen, X., Liu, B., Zhou, Y., Zhao, J., & Liu, M. (2020). Remote sensing image captioning via variational autoencoder and reinforcement learning. *Knowledge-Based Systems, 203*, Article 105920. https://doi.org/10.1016/j.knosys.2020.105920

Wang, Y., Xu, N., Liu, A. A., Li, W., & Zhang, Y. (2022). High-order interaction learning for image captioning. *IEEE Transactions on Circuits and Systems for Video Technology, 32*(7), 4417-4430. https://doi.org/10.1109/TCSVT.2021.3121062

Varma, S., & Peter, J. D. (2026). Enhanced transformer model for video caption generation. Expert Systems, 43(6), e13392.

Xian, T., Li, Z., Tang, Z., & Ma, H. (2022). Adaptive path selection for dynamic image captioning. *IEEE Transactions on Circuits and Systems for Video Technology, 32*(9), 5762-5775. https://doi.org/10.1109/TCSVT.2022.3155795

Xu, X., Wu, X., Zhang, H., & Ma, K. (2026). Medical Image Captioning with Swin Transformer and ClinicalBERT via Gated Multimodal Fusion. Applied Soft Computing, 115731.

Yan, C., Hao, Y., Li, L., Yin, J., Liu, A., Mao, Z., Chen, Z., & Gao, X. (2022). Task-adaptive attention for image captioning. *IEEE Transactions on Circuits and Systems for Video Technology, 32*(1), 43-51. https://doi.org/10.1109/TCSVT.2021.3067449