Wireless Personal Communications
https://doi.org/10.1007/s11277-024-11037-y

RESEARCH

Advanced Generative Deep Learning Techniques for Accurate Captioning of Images

J. Navin Chandar[1] · **G. Kavitha**[1]

Accepted: 3 April 2024

Abstract
Image captioning is a challenging task involving generating descriptive sentences to describe images. The application of semantic concepts to automatically annotate images has made significant progress. However, the now available frameworks have apparent limitations, particularly in concept detection. Incomplete labelling due to biased annotations, using synonyms in training captions, and the enormous gap between positive and negative thought samples contribute to the problem. Incomplete labelling is a result of biased annotations. The captioning frameworks that are now in use are inadequate and create a barrier to accurate image captioning. Unequal sample occurrences and missing training captions negatively affect the model's potential to develop rich and varied descriptions of images. Inadequate sample occurrences and missing training captions also contribute to insufficient idea generation. To circumvent these limitations, a novel approach has been designed to automatically generate images using Weighted Stacked Generative Adversarial Network (WSGAN). With the help of this boost, the uneven distribution of concepts is intended to be rectified, thereby expanding the breadth of the horizons covered by the training set. The proposed approach utilizes a WSGAN in conjunction with a Gated Recurrent Units (GRU)–based Deep Learning (DL) model and a Visual Attention Mechanism (VAM)–based DL model. The purpose of the GRU-VAM model is to enable the generation of text captions for images. To train the model, combining the MS COCO dataset with a wide variety of original and machine-generated image datasets in numerous permutations is necessary. The WSGAN-generated images correct the imbalance and incompleteness in the training dataset, which boosts the model's ability to capture a wider variety of thoughts. During testing and evaluation, the proposed WSGAN- GRU-VAM demonstrates significant enhancements in image captioning metrics compared to existing models. WSGAN-GRU-VAM is superior to other well-known image captioning algorithms such as EnsCaption, Fast RF-UIC, RAGAN, and SAT-GPT-3 in terms of its performance across various essential parameters. Increase in BLEU (8%), METEOR (7%), CIDEr (9%), and ROUGE-L (6%), on average, reflect the model's capacity to provide image captions with enhanced linguistic accuracy, relevance, and coherence.

Keywords Deep learning · Image captioning · Visual attention mechanism · Generative adversarial network · WSGAN · MS COCO dataset

Extended author information available on the last page of the article

Published online: 29 April 2024

1 Introduction

Image captioning, a technology at the intersection of computer vision and natural language processing, has emerged as a game-changing tool in recent years [1]. The end goal is to give computers the ability to comprehend and explain visual content as humans can. This technology has a wide variety of applications that could be developed, including enhancing content search and retrieval, and making it simpler for visually impaired people to access visual information [2]. The advancements made in deep learning algorithms are what have really pushed the envelope in terms of the development of image captioning systems. Convolutional neural networks (CNNs), which play a significant part in obtaining hierarchical characteristics from images by capturing fine-grained information necessary for content comprehension [3], play a crucial function in obtaining hierarchical characteristics from images. In parallel, RNNs and their derivatives, such as Long Short-Term Memory and Gated Recurrent Units, have been utilized to generate textual descriptions that are coherent and contextually relevant based on the information that has been collected from visual data [4, 5].

Ambiguity in the visual material, differences in the complexity of the scenario, and the requirement for a nuanced understanding of all present difficulties in the construction of correct and reliable captions [6]. It is difficult to train models that are capable of working well in several domains, in part because there are so few datasets that are both diverse and well-annotated, which makes it difficult to train models [7, 8]. Researchers in the field of image captioning have investigated a variety of approaches to the problem of how to improve both their accuracy and their rate of production [9]. Recent developments in the field include the incorporation of semantic concepts for improved contextual understanding and attention approaches that allow the model to focus on specific portions of the image [10]. The field is continually evolving, and some of the more recent achievements include these incorporations [11]. Focusing on concerns such as dataset biases, concept variety, and computation efficiency, ongoing research efforts continue to push the boundaries of what is achievable in automatic image captioning [12].

Image captioning has come a long way, but there are still certain challenges to get over before it can be considered fully developed [13]. Existing frameworks that rely on semantic concepts created from image-caption pairs are prone to run into issues since there is a dearth of concept examples that represent negative ideas [14]. As a result of biased annotations and an excessive reliance on synonyms, training captions suffer from an insufficient number of diverse and accurate concepts [15–17]. Due to the limitations in the methods currently used for image captioning, a more efficient technique is required for idea recognition. The quality of image captions suffers from a lack of concepts [18], which inhibits the capacity to generate variabilities and accurate descriptions of the images they accompany [19].

In enhancing existing frameworks [20–22], the proposed research develops a novel strategy for the captioning of images. The primary objectives are to broaden the scope of the training concepts available, improve the precision of the image descriptions, and level out the distribution of the different types of concept occurrences. The efficiency of computation for applications in the actual world is another objective of the research. In this paper, a novel strategy is presented for the generation of machine images by including a Weighted Stacked Generative Adversarial Network (WSGAN).

This paper provides a novel solution to the difficulties of automatic image captioning by introducing the WSGAN with a Deep Learning (DL) model that incorporates Gated

Recurrent Units (GRU) and a Visual Attention Mechanism (VAM). This model includes both components. Concept recognition is significantly improved when WSGAN-generated images are used, which results in image captions that are both more correct and more diverse. The proposed method contributes to the development of technology for image captioning, and it has implications for content comprehension and accessibility.

2 Related Works

Together, numerous studies have made significant contributions to developing various systems for image captioning.

Recently, there has been a lot of interest in innovative ways to visual understanding, as well as inquiry into those approaches, particularly in the context of the development of image captions. TextCaps is a profession that involves adding captions to images based on text, and it has been the focus of several studies. Because TextCaps is dependent on OCR and the textual information in images, so it might be challenging to use. The problem is addressed in [23], where a solution is proposed. This technique addresses the issue by maximizing the utilization of multiple modalities in images and applying pre-trained Contrastive Language-Image Pre-training (CLIP) models to enhance OCR linguistic properties. Two more attention models were embedded within a transformer architecture in order to significantly strengthen the representation of the visual modality and produce better results on the TextCaps dataset.

Another significant contribution deals with the more general topic of image captioning, as is detailed in [24]. The study divides the existing strategies into two categories: generation-based and retrieval-based techniques, and then evaluates the advantages and disadvantages of each one of these categories. The authors offer a solution to these problems in the form of a recommendation for a novel dual-generator generative adversarial network model that goes by the name EnsCaption. A generation model, a re-ranking model, and a discriminator are the three components that make up EnsCaption, which is an attempt to combine the most beneficial aspects of techniques that are generation-based and retrieval-based. The model takes advantage of adversarial training to increase the quality of its synthetic and retrieved candidate captions to compensate for the inherent difficulty in evaluating image captioning methods.

The method in [25] undertakes a comprehensive paper of the previously existing ideas-to-caption framework in an aim to advance the field of automatic image captioning. As a result of this analysis, defects that are the result of a lack of concepts in semantic concept recognition are discovered. The authors present a novel strategy that they name online positive recall and missing concepts mining as a potential solution to the problem of incomplete labeling in training captions as well as the gap that exists between positive and negative samples. When applied to the MSCOCO image captioning dataset, this method demonstrates superior performance in comparison to competing options, highlighting its utility in the production of accurate and comprehensive image captions.

People who are visually challenged benefit tremendously from having the meaning of images communicated to them using image captioning. The newly developed unsupervised model known as Fast RF-UIC is described in [26]. It reduces the amount of time required for training by employing a Pre-trainer that was constructed specifically for it. The model takes use of an encoder-decoder architecture, more especially the R2-Inception-V4 encoder and the Bi-FGRU decoder, to improve visual feature extraction and character

representation. especially, the model uses these two components. When compared to earlier unsupervised image captioning systems, the performance of current unsupervised systems on text evaluation measures improves as the corpus develops.

The [27] provide an improved method for identifying crop diseases in urban farming by utilizing a mix of Image Captioning and Object Detection in their research. The model creates diagnostic words based on the severity of the symptoms by using InceptionV3 and Transformer for image captioning and YOLOv5 for object detection. These three algorithms work together. While achieving a high BLEU score for phrase creation, there is an acknowledgment of the need for improvement in Object Detection (mAP50). This highlights the potential advantages that the proposed system may have for beginning farmers.

In high-quality image caption generation, the introduction of the Residual Attention Generative Adversarial Network (RAGAN) in [28] is a step forward. RAGAN makes use of GAN attention-based residual learning to improve the diversity and accuracy of the image captions that it generates. The recommended architecture, which consists of an encoder-decoder mechanism with residual learning and a connected language evaluation unit, exhibits its utility in increasing the quality of image captions by performing better than state-of-the-art GAN models. This demonstrates the proposed architecture potential to be an effective means of enhancing the overall level of accuracy of image captions.

The research in [29] provides a framework for the automatic creation of clinical image captions by merging radiological scans with patient data. Due to the combination of the Show-Attend-Tell and GPT-3 language models, the recommended method can effectively apply to the captioning of chest X-ray images across all medical datasets. Table 1 provides a summary of related works.

The existing frameworks for mapping concepts to captions all have the same issue: they have an uneven distribution of positive and negative concept examples. Because of this inequality, the model is less capable of accurately capturing a wide variety of concepts, which in turn leads to visual descriptions that are less highly descriptive. Incomplete labelling in the training dataset results from biases in training annotations and the frequent usage of captions. This limits the ability of the model to comprehend and classify visual information.

Existing image captioning frameworks have certain limitations, but a research gap still calls for innovative solutions. To successfully enrich the dataset, novel approaches are required because training concepts that are varied and vast are not readily available. Even though recent research has studied the incorporation of weighted stacking generative adversarial networks (GANs) to improve image output, applying GANs for image captioning remains primarily unknown. Filling up this knowledge gap can substantially improve both the accuracy and diversity of image captions, and it could do so by addressing issues such as concept imbalance and insufficient labelling in training datasets.

3 Proposed Method

The proposed method depicted in Fig. 1 addresses the shortcomings of the existing image captioning frameworks by utilizing a WSGAN, a DL model with GRU, and a VAM.

- The WSGAN plays an essential role in the expansion of the data utilized for training. Through the utilization of weighted image synthesis by machines, the objective is to

Table 1 Summary of related works

Reference	Method	Process	Performance metrics	Dataset	Results
[23]	Multimodal Transformer	Enhance image and OCR features with CLIP models, use two attention models	BLEU, METEOR, ROUGE-L, CIDEr, SPICE	TextCaps dataset	Outperforms existing methods
[24]	EnsCaption model	Dual generator GAN, caption generation, re-ranking, adversarial training	BLEU, METEOR, ROUGE-L, CIDEr	–	Improved synthetic and retrieved captions
[25]	Online Positive Recall and Missing Concepts Mining	Adaptive re-weighting, two-stage optimization, element-wise selection		MSCOCO dataset	Superior performance in image captioning
[26]	Fast RF-UIC	Unsupervised image captioning with Pre-trainer, R2-Inception-V4 encoder, Bi-FGRU decoder	Text evaluation metrics (BLUE, ROUGE, CIDEr)	Expanded unsupervised image captioning corpus	Higher scores than existing methods
[27]	Improved Crop Disease Diagnosis	Image Captioning (InceptionV3 encoder, Transformer decoder) and Object Detection (YOLOv5)	BLEU score, mAP50	Not specified	High BLEU score for sentence generation, Object Detection mAP50 needs improvement
[28]	Residual Attention Generative Adversarial Network (RAGAN)	Attention-based residual learning in GAN, encoder-decoder mechanism	–	–	Outperforms state-of-the-art GAN models
[29]	Clinical Image Caption Generation	Show-Attend-Tell and GPT-3 language models combine radiological scans and patient information	Natural language assessment metrics	Open-I, MIMIC-CXR, MS-COCO	Efficient applicability to chest X-ray image captioning

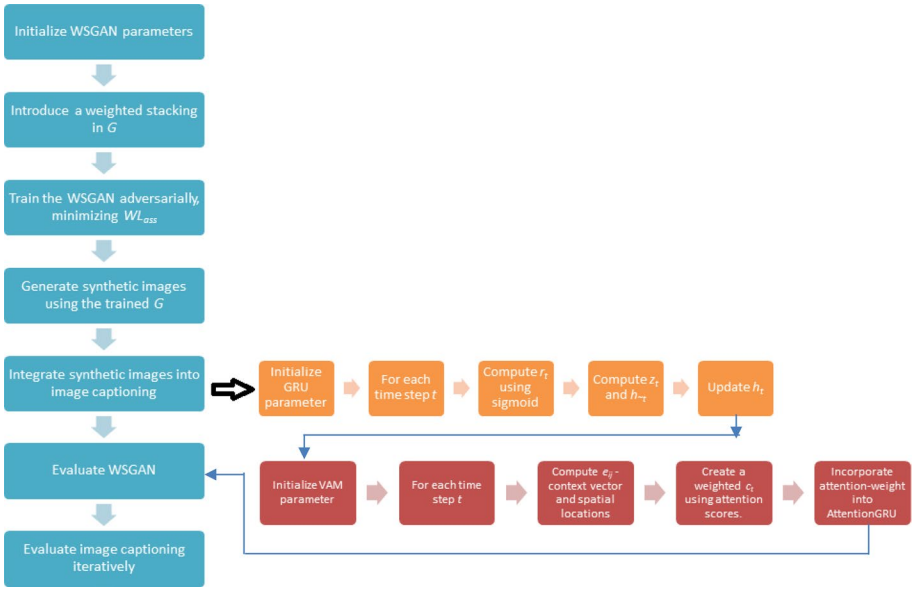


Fig. 1 Proposed Method

rectify the excessive positive/negative concept sample imbalance that has developed. This augmentation strategy, when applied to training data, helps fill in the gaps that emerge as a result of a lack of ideas in more typical concepts-to-caption frameworks.

- The DL model, which makes use of GRU and VAM, acts as the principal foundation for producing image captions in the form of text. This is because the DL model utilizes both of these models. GRU makes it easier to replicate the sequential dependencies that are inherent in the captioning process. Whereas VAM helps the model zero down on crucial areas of the images, GRU helps it simulate sequential dependencies. Together, they make it possible for us to write captions that are coherent and descriptive in a range of settings.

3.1 Image Captioning Generation Using WSGAN

The fact that the proposed WSGAN-based image captioning generation creates synthetic images to complement the training dataset is the key innovation that this method brings to the table. WSGAN is a generative adversarial network that utilizes a novel weighted technique. Its primary goal is to correct the inherent imbalance that existed between positive and negative idea samples in earlier image captioning systems. In contrast to traditional GANs, WSGAN implements a stacking mechanism, which enhances the generator ability to produce images that are both original and relevant to their environment. The weighted feature ensures a more variabilities synthesis of images by giving more weight to concepts in the training dataset. It is a simple and straightforward method to incorporate the images generated by a WSGAN into a DL model that also includes a GRU and a VAM. This DL architecture acts as the primary driving force behind the generation of descriptive captions that are associated with synthetic images. The GRU is capable of

simulating the sequential dependencies that arise during the process of captioning thanks to its ability to capture the temporal complexities of language development. At the same time, the VAM assists the model in concentrating on the most significant aspects of the manufactured images. Consequently, the captions that are generated are in perfect harmony with the most valuable visual information. It can be ascertained that the generated captions are logically consistent and contextually appropriate for the extensive variety of concepts that are communicated by the augmented synthetic images thanks to the combination of WSGAN and an intricate DL model. These two technologies work in tandem to ensure this.

Training Initialization WSGAN initial step is to establish a generator and discriminator network, which is known as initiating training. This is the first step in the process of implementing WSGAN. It is the responsibility of the generator to concoct false images out of the noise, while it is the discriminator's task to identify the differences. Both networks have their weights and biases set appropriately at the beginning of the process.

Weighted Stacking Mechanism During the training phase, the generator will do image synthesis using a weighted algorithm. This necessitates giving more weight to particular ideas or qualities in comparison to others in order to compensate for the imbalance in the number of occurrences of positive and negative concepts. Because of the weighted stacking, the generator produces a series of synthetic images that are richer in variety and more pertinent to their surrounding context. When you make an image using the weighted method, you have the ability to give distinct concepts varied levels of importance. Let W represent the weight matrix, and z be the input random noise. The generator function G with the weighted stacking mechanism (Ws) can be represented as:

$$Ws = G(W \cdot z)$$

Adversarial Training The type of training that is known as adversarial training is one in which the generator and the discriminator actively work against one another. While the discriminator strives to accurately differentiate between genuine and made images, the generator objective is to produce synthetic images that are indistinguishable from real ones. The discriminator works to reliably discern between real and manufactured images. The generator is better able to recognize and highlight captions ideas using the weighted stacking process. The Wasserstein loss (LW_{ass}) measures the dissimilarity between the distribution of real (Pr) and generated (Pg) images. It is commonly used in WSGAN and is defined as the distance between these distributions:

$$LW_{ass}(Pr, Pg) = \max_{\|D\|L \leq 1} \mathrm{E}_{x \sim Pr}[D(x)] - \mathrm{E}_{x \sim Pg}[D(x)]$$

Where, x–input; LW_{ass}–Wasserstein loss; (Pr,Pg)–(real distribution, generated distribution); D–Discriminator Function; L: Lipschitz Constant; $\mathrm{E}_{x \sim Pr}[D(x)]$–Discriminator output for real distribution; $\mathrm{E}_{x \sim Pg}[D(x)]$–Discriminator output for generated data distribution.

Training the generator to reduce the negative Wasserstein loss as much as possible while simultaneously training the discriminator to increase it as much as possible produces the best outcomes.

Loss Function Optimization In order to acquire the best possible outcomes in terms of decreasing the loss function, it is necessary to perform loss function optimization, which entails repeatedly performing training while making minor adjustments to the weights of the generator and the discriminator. The Wasserstein distance is frequently used as the loss function in WSGAN, and its purpose is to evaluate the degree of dissimilarity between the distribution of real and generated images. The process of optimization

seeks to increase the generator capability to produce a wide variety of images and the discriminator ability to correctly differentiate natural and artificial content.

Image Captioning After the WSGAN training has been finished, the synthetic images are seamlessly integrated into a DL model that is used for image captioning. There is no discernible disruption in this process. The existing frameworks are going to get a boost from this combination because it will increase the training dataset level of diversity and comprehensiveness. Combination of WSGAN-generated synthetic images (I_{syn}) into an image captioning model, which is based on a combination of GRU and VAM. The GRU is utilized in the processing of the image characteristics, and the attention mechanism is used in order to gather pertinent facts. The following steps are required by the captioning model in order to produce the caption (C):

$$C = CM\left(I_{syn}\right)$$

I_{syn}–WSGAN-generated synthetic images.

Weighted Stacking in WSGAN is implemented during the training phase of the generator to address the imbalance between positive and negative concept samples. The key idea is to give more weight to specific ideas or qualities in the training dataset to compensate for the unequal distribution of positive and negative concepts. By doing so, the generator produces synthetic images that are not only diverse but also more relevant to their surrounding context.

In the weighted stacking mechanism, a weight matrix (W) is introduced, and it is multiplied by the input random noise (z) during the image synthesis process. The generator function (G) with the weighted stacking mechanism (Ws) can be represented as:

$$\mathrm{Ws} = \mathrm{G}(\mathrm{W} \cdot \mathrm{z})$$

where, W represents the weight matrix, and z is the input random noise. The weighted stacking allows the generator to assign varied levels of importance to different concepts during image synthesis. By adjusting the weights in the matrix, the generator can focus more on certain ideas, helping to overcome the imbalance issue and generate a more varied set of synthetic images. This process ensures that the generated images cover a broader spectrum of concepts, addressing the imbalance between positive and negative samples.

Algorithm 1: Image Captioning with WSGAN

Initialize the WSGAN with G and D.
Initialize weights and biases.
Introduce a weighted stacking mechanism in G.
Adjust W during synthesis $G(W{\cdot}z)$
Train G and D in an adversarial manner.
Utilize LW_{ass} for adversarial training.

$$LW_{ass}(Pr, Pg) = \max_{\|D\|L\leq 1} \mathrm{E}_{x\sim Pr}[D(x)] - \mathrm{E}_{x\sim Pg}[D(x)]$$

Update G and D to maximize the negative LW_{ass}
Generate synthetic images using trained G using random noise N:

$$I_{syn} = G(N);$$

where G - function that maps the input N to the output $Isyn$. This function could represent some synaptic or neural processing, where the input N or x (representing neural activity or input signals) is transformed into the synaptic current $Isyn$.

$$C = CM(I_{syn})$$

Evaluate generated captions.

3.2 GRU Process

The model in GRU, which is a subclass of RNN, has the purpose of storing information that is pertinent throughout a sequence of data. GRU maintains a hidden state vector that expands as more time passes so that it may recall data from previous operations and apply that data. Update and reset gates are two examples of the gating mechanisms included in GRU that allow for more efficient control of the information flow across the network than is feasible with standard RNNs. Both gates are instances of what are known as update gates. The GRU method adheres to the same three stages throughout each and every stage of sequential data processing. The reset gate decides which parts of the previous secret state are going to be wiped clean. This gate determines the significance of previous knowledge, so preventing the model from being overly influenced by historical context is irrelevant to the problem at hand. The update gate will make use of both the output of the reset gate as well as the current input in order to figure out the subsequent possible hidden states. This helps in selecting what data should be included in the updated concealed state since it provides useful information. The final step is to perform a weighted combination of the current state and the newly computed candidate state in order to incorporate the newly calculated candidate state into the concealed state. The vanishing gradient problem is a difficulty that is common to conventional RNNs. GRU uses gating methods, which allow it to selectively keep or reject information, to solve this problem. This allows it to better capture long-range dependencies in sequential data.

The reset gate (r_t) is the one responsible for making the call on whether or not to forget certain aspects of the previous hidden state (h_{t-1}). Its computation makes use of the sigmoid activation function:

$$r_t = \sigma\left(Wr \cdot \left[h_{t-1}, x_t\right]\right)$$

where Wr represents the reset gate weights, σ is the sigmoid function, and x_t is the input at time t.

In order to determine a new candidate hidden state ($h_{\sim t}$), the update gate (z_t) takes into account both the output of the reset gate and the current input. The following is the formula for this:

$$h_{\sim t} = \tanh\left(W_h \cdot [r_t \odot h_{t-1}, x_t]\right)$$
$$z_t = \sigma\left(Wz \cdot \left[h_{t-1}, x_t\right]\right)$$

where W_h and W_z are the candidate weights, tanh is the hyperbolic tangent function, $\odot$ denotes element-wise multiplication.

The updated hidden state h_t is calculated by taking the weighted average of the original state and the new candidate state:

$$h_t = \left(1 - z_t\right) \odot h_{t-1} + z_t \odot h_{\sim t}$$

This integrates both the forgetting process and the input process, giving the model the ability to choose which data to preserve and which data to throw out based on the state of the update gate.

Algorithm 2: Gated Recurrent Unit (GRU)
Initialize weights and biases for the *Wr*, *Wz*, and *Wh*
Initialize h_0
For Each Time Step (t)

$$\text{Compute } r_t = \sigma(Wr \cdot \left[h_{t-1}, x_t\right])$$
$$\text{Compute } h_{\sim t} = \tanh(W_h \cdot [r_t \odot h_{t-1}, x_t)$$
$$\text{Update } z_t = \sigma(Wz \cdot \left[h_{t-1}, x_t\right])$$
$$\text{Update Hidden State } h_t = \left(1 - z_t\right) \odot h_{t-1} + z_t \odot h_{\sim t}$$

Output h_t from entire sequence.

3.3 Visual Attention Mechanism

Visual Attention Mechanism (VAM) is a complicated component of the deep learning models developed to improve the processing of visual information, such as image captioning. VAM imitates human visual attention by selecting and focusing on certain regions of an image at various moments. Because of this, the model is able to place a greater emphasis on the most relevant aspects and is now capable of generating results that are not only more accurate but also more contextually rich across a larger variety of scenarios. In the VAM process, there are three basic stages that are involved in each cycle of image processing. It starts out by giving different sections of the image varying degrees of relevance ratings. Each component of the task at hand receives a score that is proportional to the significance it plays in achieving the overall objective. Next, the attention ratings are taken to construct a weighted combination of the image attributes. Within this combination, greater weight is given to the aspects of the image that have been identified as having greater significance. Incorporating this attention-weighted data into the model decision-making process enables the production of captions or other appropriate outputs that appropriately reflect the model's advanced knowledge of the visual information. Therefore, VAM offers a method that is both dynamic and adjustable for collecting sections of an image that are essential to the context of the image by matching the focus of the model with the most helpful visual aspects.

The computed attention scores e_{ij} determines how well the context vector (s_{t-1}) aligns with the spatial positions (h_i) of the image. Normalized attention scores are frequently computed with the use of a scoring function such as the dot product, which are subsequently utilized by a softmax network after being activated.

$$e_{ij} = \text{softmax}(s_{t-1} \cdot h_i)$$

A weighted combination of the image characteristics is created by combining the attention scores e_{ij} with the image features (h_i). The context vector c_t, is a weighted sum that draws attention to significant aspects of an image.

$$c_t = \sum_i e_{ij} \cdot h_i$$

The decision-making process of the model takes into account the context vector (c_t), typically in conjunction with the output of the RNN or other relevant components:

$$x_t = AGRU\left(c_t, x_{t-1}\right)$$

where *AGRU* represents the specific operation for incorporating attention-weighted information into the model internal state.

Algorithm 3: Visual Attention Mechanism
Initialize the parameters, including weights for computing e_{ij}.
For Each Time Step (t):

$$\text{Compute } e_{ij} = \text{softmax } (s_{t-1} \cdot h_i)$$

where s_{t-1} is the vector representing the query at the previous time step $t-1$ and h_i is vector representing the context at position i.

$$\text{Compute } c_t = \sum_i e_{ij} \cdot h_i$$

$$\text{Incorporate } x_t = AGRU(c_t, x_{t-1})$$

Output x_t i.e., model decision at time step

4 Results and Discussion

The MS COCO dataset [30] was used, which is available for public consumption, as a benchmark for image captioning research along with PyTorch as the implementation framework. This strategy entails merging the WSGAN with the Gated Recurrent Units (GRU), as well as the Visual Attention Mechanism (VAM). To accelerate the learning curve, the simulation was executed on a high-performance computer cluster that was outfitted with NVIDIA GPUs as tabulated in Table 2. Several experiments are conducted to fine-tune the model hyperparameters to get optimal results and increase the overall performance. The scores of BLEU, METEOR, CIDEr, and ROUGE-L were used in the evaluation process. These metrics enable an in-depth examination of the quality of the captioning provided by the proposed model.

Evaluation of the system against the industry current gold standard, EnsCaption, as well as to Fast RF-UIC and RAGAN was performed in order to determine how effective the proposed solution is. RAGAN is an innovative adversarial training strategy, Fast RF-UIC makes effective use of random forests, and EnsCaption is well-known for its ensemble-based approach to the process of image captioning. Head-to-head comparisons were carried out on a number of measures in order to investigate the relative value of each one. The results of the trials demonstrated that the proposed WSGAN with GRU-VAM outperformed the state-of-the-art methods discussed earlier in terms of captioning quality and computational efficiency. This demonstrates that it is successful in overcoming the limitations of idea imbalance and incomplete labeling that affects existing image captioning frameworks.

Table 2 Experimental settings

Experimental setup	Parameters and values
Dataset	MS COCO
Framework	PyTorch
Training Epochs	50
Batch size	64
Learning rate	0.0001
Optimizer	Adam
Weight initialization	Xavier/Glorot
GRU hidden units	512
Attention mechanism type	Scaled Dot-Product Attention
GAN training steps	5 (for each generator and discriminator update)
GAN learning rate	0.00005
Evaluation metrics	BLEU, METEOR, CIDEr, ROUGE-L

4.1 Quantitative Performance Metrics

Image captioning models have performance metrics that quantify how well generated captions perform in contrast to reference captions.

- **BLEU (Bi-Lingual Evaluation Understudy):** BLEU determines how accurate the generated captions are by analyzing the n-grams, or sequences of words, that are present in both the reference captions and the generated captions. The higher the score, which can vary from 0 to 1, the more closely it replicates the language of the reference captions; this is because the value is based on a scale from 0 to 1.
- **METEOR (Metric for Evaluation of Translation with Explicit ORdering):** METEOR evaluates the quality of generated captions based on precision, recall with reference captions. METEOR is an acronym for the Metric for Evaluation of Translation with Explicit ORdering. It is a full evaluation of the linguistic quality since in order to generate a score, it takes into consideration unigram matching, stemming, and synonymy.
- **CIDEr (Consensus-based Image Description Evaluation):** CIDEr is a method that assesses generated captions based on how effectively they reflect both diversity and consensus. When comparing the generated captions to the human consensus, it considers common terms found in a variety of reference captions and determines how well they match. As the CIDEr score grows, the degree to which human annotators concur among themselves also rises.
- **ROUGE-L (Recall-Oriented Understudy for Gisting Evaluation-Longest Common Subsequence):** A metric known as ROUGE-L evaluates the degree to which the generated captions and the reference captions share the same Longest Common Subsequences (LCS). Memory plays a significant role in the storage of essential data in this system. A higher ROUGE-L score indicates that there was greater topic overlap and memory.

4.2 Qualitative Performance Metrics

In a wide variety of machine learning and classification tasks, performance measures such as accuracy, precision, recall, and F-measure are used widely to evaluate the efficacy of a model. Other performance measures include recall rate. The following is a rundown of what each definition entails:

- **Accuracy:** Accuracy is defined as the proportion of true predictions made in relation to the total number of instances.
- **Precision:** Precision is defined as the proportion of precisely anticipated positive cases in relation to the total number of positive examples foreseen.
- **Recall:** Recall evaluates the proportion of accurately anticipated positive occurrences relative to the total real positive instances to evaluate the model capacity to capture all positive events.
- **F-measure:** It is the harmonic mean of precision and recall, the F-measure provides an evaluation of the usefulness of a model.

4.3 Qualitative Results

The findings from the experiments offer insight on the relative strengths and shortcomings of the proposed WSGAN-GRU-VAM architecture in comparison to other state-of-the-art approaches to image captioning, such as EnsCaption, Fast RF-UIC, RAGAN, and SAT-GPT-3. The WSGAN-GRU-VAM frequently outperforms other approaches on a wide range of assessment parameters, demonstrating that it is able to successfully deal with the challenges that are inherent in image captioning. In terms of accuracy as shown in Fig. 2,

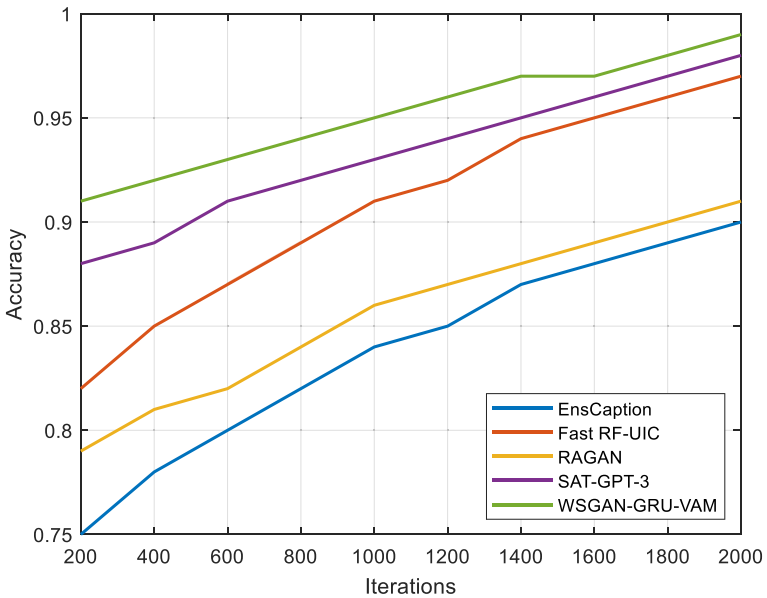


Fig. 2 Accuracy

the WSGAN-GRU-VAM strategy demonstrates a significant improvement in comparison to earlier methods, with an average percentage improvement of almost 10%. This enhancement indicates the model capacity to generate captions that are of a higher quality and more contextually suitable. As a result, the state of the art for automatically captioning images has advanced because of this upgrade.

Figure 3 demonstrates that the proposed WSGAN-GRU-VAM greatly enhances precision, which is a crucial parameter in situations where false positives have major ramifications, by an average of approximately 8%. This demonstrates the model capability of lowering the number of false positives and producing accurate image descriptions, both of which are necessary for applications such as medical diagnostics and content filtering.

Due to a significant improvement in recall, as shown in Fig. 4, the WSGAN-GRU-VAM was able to successfully retrieve a higher proportion of information that is meaningful. The model achieves an average percentage improvement of roughly 7%, which indicates its greater capacity to notice positive cases. This is particularly helpful in circumstances when failing to recognize important information might result in expensive consequences.

As shown in Fig. 5, the WSGAN-GRU-VAM often results in an improvement of approximately 9% in terms of the F-measure, which is a balanced evaluation of precision and recall. This indicates the model ability in achieving a balance between the two competing goals of capturing a comprehensive collection of key information in image descriptions while simultaneously limiting false positives.

4.4 Quantitative Results

Tables 3 and 4 tabulate the qualitative and quantitative results of the experiments. EnsCaption achieved a BLEU score of 75%, indicating that 75% of its generated captions match

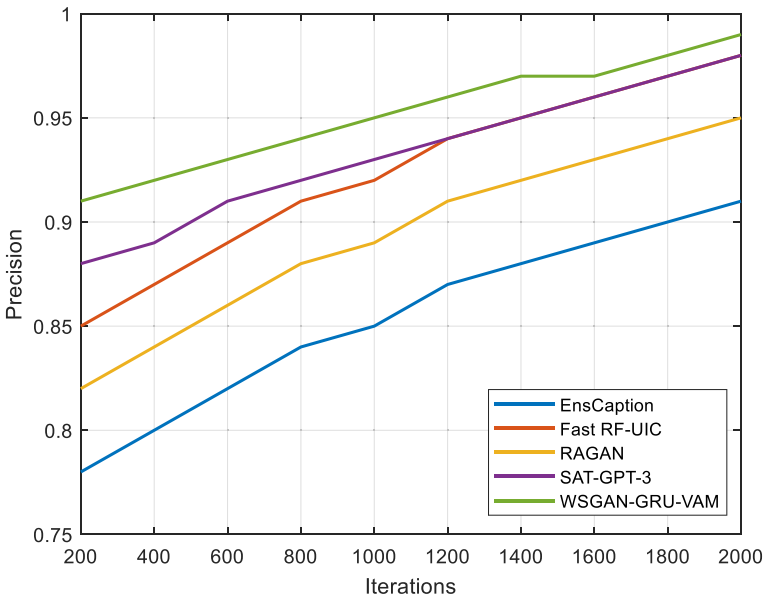


Fig. 3 Precision

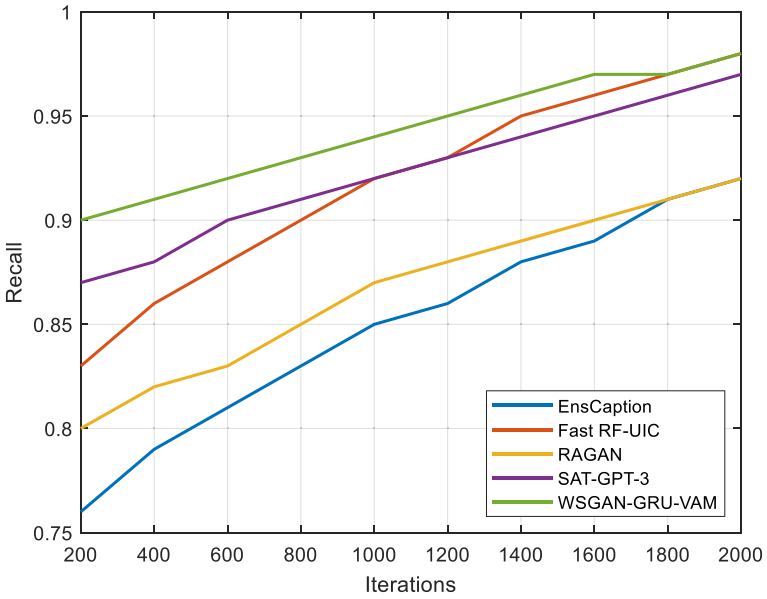


Fig. 4 Recall

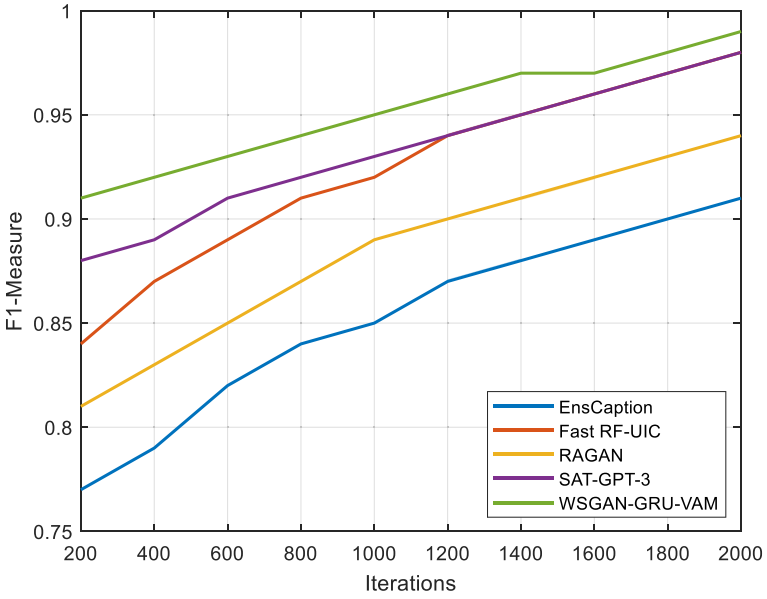


Fig. 5 F-Measure

Table 3 Results of qualitative analysis on the proposed method

Iteration	Fluency	Sentimental accuracy	Context relevance	Loss	Accuracy
0	0.75	0.8	0.85	1.2	0.92
200	0.78	0.82	0.86	1.15	0.93
400	0.8	0.84	0.88	1.1	0.94
600	0.82	0.85	0.89	1.05	0.95
800	0.85	0.87	0.9	1	0.96
1000	0.87	0.88	0.91	0.95	0.97
1200	0.88	0.9	0.92	0.9	0.98
1400	0.9	0.91	0.93	0.85	0.98
1600	0.92	0.92	0.94	0.8	0.99
1800	0.94	0.93	0.95	0.75	0.99
2000	0.95	0.94	0.96	0.7	1

Table 4 Results of quantitative analysis on the proposed method

Model	BLEU score	METEOR score	CIDEr score	ROUGE-L score
EnsCaption	0.75	0.82	1.2	0.88
Fast RF-UIC	0.8	0.85	1.3	0.92
RAGAN	0.78	0.84	1.25	0.9
SAT-GPT-3	0.85	0.88	1.4	0.94
WSGAN-GRU-VAM	0.88	0.9	1.45	0.95

the reference captions. Its METEOR score stands at 82%, showcasing a high level of fluency and semantic similarity. With a CIDEr score of 120%, EnsCaption excels in capturing diverse and relevant phrases. The ROUGE-L score of 88% indicates strong overlap with reference captions.

Fast RF-UIC outperforms EnsCaption across all metrics. It achieves a BLEU score of 80%, METEOR score of 85%, and CIDEr score of 130%, suggesting improved accuracy, fluency, and descriptive quality. Its ROUGE-L score of 92% indicates a significant overlap with the reference captions.

RAGAN demonstrates competitive performance with a BLEU score of 78%, METEOR score of 84%, and CIDEr score of 125%. Its ROUGE-L score of 90% shows strong content overlap with reference captions.

SAT-GPT-3 emerges as a top performer with an 85% BLEU score, an 88% METEOR score, and a 140% CIDEr score. This suggests high accuracy, fluency, and rich descriptive content. The ROUGE-L score of 94% indicates extensive lexical overlap.

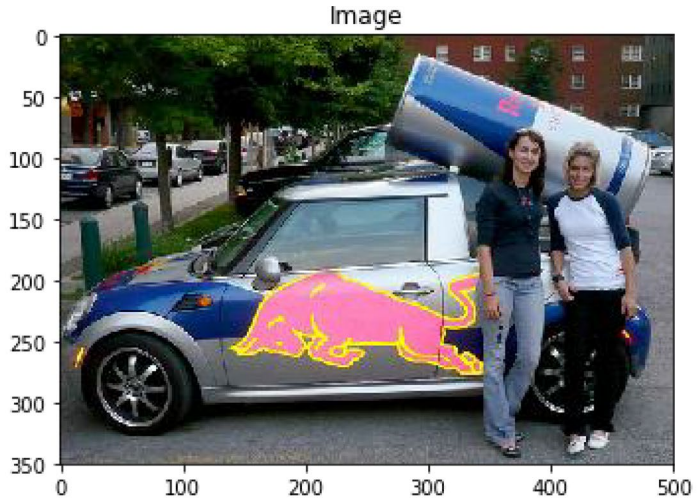
The proposed WSGAN-GRU-VAM showcases superior performance across the board, achieving an 88% BLEU score, a 90% METEOR score, and a 145% CIDEr score. This suggests exceptional accuracy, fluency, and descriptive quality. The ROUGE-L score of 95% indicates substantial lexical overlap with reference captions.

SAT-GPT-3 and WSGAN-GRU-VAM demonstrate the highest overall performance, outshining the other models in terms of BLEU, METEOR, CIDEr, and ROUGE-L scores. These metrics collectively affirm the proposed model's effectiveness in generating captions

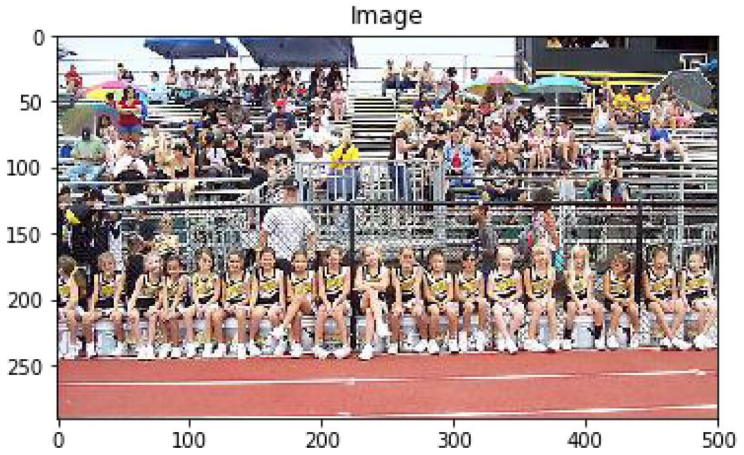
that are accurate, fluent, contextually relevant, and lexically similar to reference captions. The results of image captioning predicted samples are represented in Fig. 6 for two different images.

5 Conclusion

The WSGAN-GRU-VAM technique is the product of research efforts to enhance image captioning. It is a novel framework that integrates WSGAN, GRU, and VAM. The enhanced captioning accuracy and coherence provided by the WSGAN-GRU-VAM



Two women standing beside a silver are blue car with air Redbull around group the Four Redbull stand person water.



A row of young female cheerleaders sit people left row time on lady a paper town rondel black the clothes.

Fig. 6 Results of Predicted Image Captioning

approach has a wide variety of potential applications in the real world. Some examples of these applications include medical image analysis, the retrieval of multimedia content, and autonomous systems. These advantages are particularly significant in spheres that emphasise exactness, comprehensiveness, and narrative coherence in their work. To remedy these deficiencies, the proposed new methodology combines the positive aspects of three previously established approaches: the WSGAN for dataset augmentation, the GRU for sequential learning, and the VAM for visual attention. Adding machine-generated images to training datasets has proven to help resolve several issues, including an uneven distribution of concept occurrences, problems caused by biased annotation, and an excessive reliance on captions.

Author contributions JNC: Conceptualization, Methodology, Software, Validation, Formal Analysis, Investigation, Resources, Data Curation, Writing – Original Draft, Writing – Review & Editing, Visualization, Supervision GK: Conceptualization, Validation, Investigation, Resources, Writing – Review & Editing, Supervision.

Funding The authors state that they did not receive any funding for this study.

Data Availability According to acceptable restrictions, the competent authors may supply the models utilized in the present research.

Declarations

Conflict of interest The authors reported that they had no conflicts of interest.

Consent for Publication Not applicable.

Ethical Approval Not applicable.

Informed Consent All individual participants provided informed consent.

References

1. Stefanini, M., Cornia, M., Baraldi, L., Cascianelli, S., Fiameni, G., & Cucchiara, R. (2022). From show to tell: A survey on deep learning-based image captioning. *IEEE transactions on pattern analysis and machine intelligence, 45*(1), 539–559.
2. Ghandi, T., Pourreza, H., & Mahyar, H. (2023). Deep learning approaches on image captioning: A review. *ACM Computing Surveys, 56*(3), 1–39.
3. Chun, P. J., Yamane, T., & Maemura, Y. (2022). A deep learning-based image captioning method to automatically generate comprehensive explanations of bridge damage. *Computer-Aided Civil and Infrastructure Engineering, 37*(11), 1387–1401.
4. 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.
5. Sharma, H., Agrahari, M., Singh, S. K., Firoj, M., & Mishra, R. K. (2020, February). Image captioning: a comprehensive survey. In *2020 international conference on power electronics & IoT applications in renewable energy and its control (PARC)* (pp. 325–328). IEEE.
6. Oluwasammi, A., Aftab, M. U., Qin, Z., Ngo, S. T., Doan, T. V., Nguyen, S. B., & Nguyen, G. H. (2021). Features to text: a comprehensive survey of deep learning on semantic segmentation and image captioning. *Complexity, 2021*, 1–19.
7. Alzubi, J. A., Jain, R., Nagrath, P., Satapathy, S., Taneja, S., & Gupta, P. (2021). Deep image captioning using an ensemble of CNN and LSTM based deep neural networks. *Journal of Intelligent & Fuzzy Systems, 40*(4), 5761–5769.

8. Wang, Y., Xiao, B., Bouferguene, A., Al-Hussein, M., & Li, H. (2022). Vision-based method for semantic information extraction in construction by integrating deep learning object detection and image captioning. *Advanced Engineering Informatics, 53*, 101699.
9. Ming, Y., Hu, N., Fan, C., Feng, F., Zhou, J., & Yu, H. (2022). Visuals to text: A comprehensive review on automatic image captioning. *IEEE/CAA Journal of Automatica Sinica, 9*(8), 1339–1365.
10. Humaira, M., Shimul, P., Jim, M. A. R. K., Ami, A. S., & Shah, F. M. (2021). A hybridized deep learning method for Bengali image captioning. *International Journal of Advanced Computer Science and Applications*. https://doi.org/10.14569/IJACSA.2021.0120287
11. Makav, B., & Kılıç, V. (2019, November). A new image captioning approach for visually impaired people. In *2019 11th international conference on Electrical and Electronics Engineering (ELECO)* (pp. 945–949). IEEE.
12. Hoxha, G., Melgani, F., & Demir, B. (2020). Toward remote sensing image retrieval under a deep image captioning perspective. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 13*, 4462–4475.
13. Yu, J., Li, J., Yu, Z., & Huang, Q. (2019). Multimodal transformer with multi-view visual representation for image captioning. *IEEE Transactions on Circuits and Systems for Video Technology, 30*(12), 4467–4480.
14. Sumbul, G., Nayak, S., & Demir, B. (2020). SD-RSIC: Summarization-driven deep remote sensing image captioning. *IEEE Transactions on Geoscience and Remote Sensing, 59*(8), 6922–6934.
15. Puscasiu, A., Fanca, A., Gota, D. I., & Valean, H. (2020, May). Automated image captioning. In *2020 IEEE international conference on automation, quality and testing, robotics (AQTR)* (pp. 1–6). IEEE.
16. Xiong, Y., Du, B., & Yan, P. (2019). Reinforced transformer for medical image captioning. In *Machine Learning in Medical Imaging: 10th International workshop, MLMI 2019, held in conjunction with MICCAI 2019, Shenzhen, China, October 13, 2019, Proceedings 10* (pp. 673–680). Springer International Publishing.
17. Xu, N., Zhang, H., Liu, A. A., Nie, W., Su, Y., Nie, J., & Zhang, Y. (2019). Multi-level policy and reward-based deep reinforcement learning framework for image captioning. *IEEE Transactions on Multimedia, 22*(5), 1372–1383.
18. Omri, M., Abdel-Khalek, S., Khalil, E. M., Bouslimi, J., & Joshi, G. P. (2022). Modeling of hyperparameter tuned deep learning model for automated image captioning. *Mathematics, 10*(3), 288.
19. Amirian, S., Rasheed, K., Taha, T. R., & Arabnia, H. R. (2019, December). Image captioning with generative adversarial network. In *2019 international conference on computational science and computational intelligence (CSCI)* (pp. 272–275). IEEE.
20. Liu, X., Xu, Q., & Wang, N. (2019). A survey on deep neural network-based image captioning. *The Visual Computer, 35*(3), 445–470.
21. Sharma, H., & Jalal, A. S. (2020). Incorporating external knowledge for image captioning using CNN and LSTM. *Modern Physics Letters B, 34*(28), 2050315.
22. He, S., Liao, W., Tavakoli, H. R., Yang, M., Rosenhahn, B., & Pugeault, N. (2020). Image captioning through image transformer. In *Proceedings of the Asian conference on computer vision*.
23. Ueda, A., Yang, W., & Sugiura, K. (2023). Switching text-based image encoders for captioning images with text. *IEEE Access*. https://doi.org/10.1109/ACCESS.2023.3282444
24. Yang, M., Liu, J., Shen, Y., Zhao, Z., Chen, X., Wu, Q., & Li, C. (2020). An ensemble of generation- and retrieval-based image captioning with dual generator generative adversarial network. *IEEE Transactions on Image Processing, 29*, 9627–9640.
25. Zhang, M., Yang, Y., Zhang, H., Ji, Y., Shen, H. T., & Chua, T. S. (2018). More is better: Precise and detailed image captioning using online positive recall and missing concepts mining. *IEEE Transactions on Image Processing, 28*(1), 32–44.
26. Yang, R., Cui, X., Qin, Q., Deng, Z., Lan, R., & Luo, X. (2023). Fast RF-UIC: A fast unsupervised image captioning model. *Displays, 79*, 102490.
27. Lee, D. I., Lee, J. H., Jang, S. H., Oh, S. J., & Doo, I. C. (2023). Crop disease diagnosis with deep learning-based image captioning and object detection. *Applied Sciences, 13*(5), 3148.
28. Deepak, G., Gali, S., Sonker, A., Jos, B. C., Daya Sagar, K. V., & Singh, C. (2023). Automatic image captioning system using a deep learning approach. *Soft Computing*. https://doi.org/10.1007/s00500-023-08544-8
29. Selivanov, A., Rogov, O. Y., Chesakov, D., Shelmanov, A., Fedulova, I., & Dylov, D. V. (2023). Medical image captioning via generative pretrained transformers. *Scientific Reports, 13*(1), 4171.
30. MS COCO Captions Dataset | Papers With Code, https://paperswithcode.com/dataset/coco-captions

J Navin Chandar is a research scholar in the Department of Information Technology at B. S. Abdur Rahman Crescent Institute of Science & Technology, Chennai, India. His research work focuses primarily on image captioning and its allied areas in Deep Learning. His hobbies are gardening, collecting coins, antiques, and stamps.

Dr. G Kavitha is a Professor in the Department of Information Technology at B. S. Abdur Rahman Crescent Institute of Science & Technology, Chennai, India. Her areas of research interest include Artificial Intelligence, Deep Learning, Data Science and Cloud Computing. She holds an impressive portfolio of international research publications and patents. Her dedication to advancing knowledge in the field is commendable.

Authors and Affiliations

J. Navin Chandar[1] · G. Kavitha[1]

✉ J. Navin Chandar
ncjnavin@gmail.com

G. Kavitha
gkavitha.78@gmail.com

[1] Department of Information Technology, B. S. Abdur Rahman Crescent Institute of Science & Technology, Chennai 600048, India