Annals of Emerging Technologies in Computing (AETiC)
Vol. 9, No. 2, 2025

Research Article

Improving Image Captioning Accuracy Using Advanced Deep Learning Techniques

Navin Chandar Jacob, Kavitha Ganesh* and Aakash Sethuraman

B. S. Abdur Rahman Crescent Institute of Science & Technology, India
ncjnavin@gmail.com; gkavitha.78@gmail.com; aakashsethu01@gmail.com
*Correspondence: gkavitha.78@gmail.com

Received: 19th November 2023; Accepted: 27th March 2025; Published: 1st April 2025;

Abstract: Image Captioning is a widely used and impactful application of Deep Learning that involves describing an image concisely and accurately. Researchers have adopted various strategies to build systems that are efficient to use in a wide range of real-life applications. The key challenges encountered are twofold - first, the need for a large volume of human created images and their corresponding captions and second, computationally intensive training required to build the model. To tackle both the challenges effectively, a novel architecture called Stacked GAN and Gated Recurrent Units Image Caption generator (STAGRIC) is proposed to accomplish the two objectives. The novelty in the architecture addresses the design concerns of building an efficient and accurate model with limited data. The first objective is accomplished using stacked GAN to synthesise images from captions which are used to augment the datasets for training. This approach supports the generation of an accurate model with limited availability of original data. The second objective, to build a model that is computationally less intensive, is accomplished using GRU based visual attention mechanism to generate captions from images. The proposed STAGRIC model is tested using MS COCO dataset and the model evaluation is performed using different combinations of images and captions datasets. The evaluation results demonstrated improved image captioning analysis metrics, and the BLEU-1 scores increased to above 75% which is higher than similar models in this space. Prospective techniques to further improve the model performance to produce higher evaluation scores are discussed in the concluding section.

Keywords: ***Deep Learning; Gated Recurrent Units; Generative Models; Image Captioning; Image Synthesis; Recurrent Neural Networks; Stacked Generative Adversarial Network***

1. Introduction

Deep Learning has opened a whole new segment of applications, primarily in the areas of Computer Vision, Language modelling and many more [1]. Today's systems are much more reliable in detecting faces or voices and fakes as well [2]. A wide variety of revolutionary architectures in this space with promising results has garnished attention from a large research community that has taken this field forward by leaps and bounds [3-4]. One crucial application that brought Deep Learning much closer in decoding human intelligence is Image captioning [5].

Image captioning can be defined as the process of generating human readable text description for an image [6-8]. Convolutional Neural Network is used to detect the features of the image elements. Next, a modified Recurrent Neural Network handles sequential data well to perform language modelling. Image captioning falls under the category of sequence-to-sequence model [9]. The model converts images which can be assumed as a sequence of pixels to a sequence of words. Most Image captioning models use the encoder decoder framework to accomplish their task [10].

Navin Chandar Jacob, Kavitha Ganesh and Aakash Sethuraman "Improving Image Captioning Accuracy Using Advanced Deep Learning Techniques", *Annals of Emerging Technologies in Computing (AETiC)*, Print ISSN: 2516-0281, Online ISSN: 2516-029X, pp. 53-65, Vol. 9, No. 2, 1st April 2025, Published by International Association for Educators and Researchers (IAER), DOI: 10.33166/AETiC.2025.02.004, Available: http://aetic.theiaer.org/archive/v9/v9n2/p4.html.

The key challenge of producing efficient models is often limited by the non-availability of large datasets. A new age solution to the problem is addressed by generating synthetic data using generative algorithms. This augments the training dataset and helps in producing efficient Deep Learning models [11]. Gartner, a leading consulting firm has predicted that by 2030 synthetic data will completely dominate real data that is used for building Machine Learning models. There are two ways of generating data. The first one is called augmentation, a technique in which new data is added to the existing real-world data. For example, given the original image, a new image is generated by rotating the image or by adding filters. Augmentation is not synthetic data but aids in creating synthetic data. The second method uses generative AI models for generating synthetic images.

Generative AI uses Machine learning and deep learning algorithms to generate artificial images or text or audio or video content [12]. Techniques like Generative Adversarial Networks (GAN), Transformers and Auto encoders fall under the class of Generative Artificial Intelligence. GAN are categorised as generative models comprising two competing neural networks trained in alternate cycles. GANs are even capable to generate new images by combining two images [13]. Such manufactured images are used in this work to supplement the primary dataset for image captioning.

Machine Learning and Deep Learning are data centric, and their performance purely depends on the size of the data and its quality. Limitations in sourcing the unbiased input data will affect the quality of the model it generates [14]. The traditional process is to collect, pre-process and annotate the data. In the case of Image captioning, the two inputs are the image and its corresponding caption. The challenge arises with getting actual data and then labelling or annotating it. These models need a huge set of training data for good performance. Collecting actual images and then assigning their respective captions is time consuming and not viable.

The key research gaps include devising an efficient solution to address the unavailability of large-scale quality data. The computational complexity of the training phase is also equally challenging and needs to be addressed. The objective of this research is to architect an efficient solution to synthesise datasets by augmenting the existing dataset and to formulate an algorithm to generate the model with lower computational complexity. In this paper, a novel architecture is proposed that leverages synthetic data generated from Stacked Generative Adversarial Networks. The Stacked GAN takes text as input and generates high quality synthetic images. The model is evaluated using MS COCO dataset. Using the pre-trained image recognition model Inceptionv3, the features of the images are extracted. Gated Recurrent Units (GRU) based Visual Attention Mechanism is used for generating text captions of images and the accuracy of the generated captions is evaluated using BLEU scores.

The key contributions in this work are summarised below,

- Stacked GAN in STAGRIC architecture addresses the dataset unavailability problem by generating synthetic data.
- The image recognition pretrained model Inceptionv3 was deployed to extract the image features accurately.
- Gated Recurrent Units based visual attention mechanism was implemented in the STAGRIC architecture with reduced computational complexity.
- The work demonstrated significant improvement in the BLEU scores of the text captions generated using STAGRIC architecture.

In the subsequent sections the work is organised as follows: section 2 gives an in-depth analysis of similar works in Image Captioning and discusses the challenges in the existing models; section 3 highlights the design of proposed STAGRIC methodology; section 4 presents the outcome of the implementation results using the MS COCO dataset; section 5 summarises the conclusion and highlights potential areas for future work.

2. Literature Review

Generative Adversarial Networks (GAN) are effectively used in generating synthetic datasets for image captioning. Due to the limited diversity in real image data sets, deep neural network based architectures have been significantly used to generate synthetic images [15]. Existing works in image captioning are classified based on CNN-RNN models, Transformer based models and Graph based models.

In CNN-RNN based image captioning systems, the visual features of the image are encoded using CNN and RNN generates captions word by word based on the extracted image features [16]. The encoder-decoder frameworks need to maintain the context of the words that frame the sentence for captions. To improve image captioning, synthesis of images from text was proposed by Hossain *et al.* [17]. The authors used ground truth captions of current labelled data sets for generating synthetic data and captioning the images. GAN was used to synthesise data and Convolutional Neural Network for extracting image features. The attention GAN mechanism was employed to augment the training dataset, and the generated captions are based on available ground truth which may lead to different interpretations and biases. The captions of synthetic images were based on existing textual descriptions of labelled real images which could not capture complex visual semantics. Any relevant parts of the image may be discarded and lead to reduced efficacy of generated captions for diverse multi-context synthetic data sets.

Image caption generation using RNN models with embedded intelligence was discussed by Verma *et al.* [18]. A CNN model is employed to extract salient features from images and fed into LSTM based RNN which generates coherent captions. LSTM uses sequence prediction and predicts the next word based on the context in the sentence. The model is validated using flicker8k dataset with 8000 images which is suitable for initial experiments but is relatively small compared to larger datasets like MS COCO. The LSTM model can further be improved by integrating attention mechanism which allow the model to focus on specific parts of image and thereby generate more accurate captions. Further, LSTM uses three gates leading to higher computational complexity and this must be reduced for efficiently training the models at a large scale.

Attention-based image captioning models improve upon traditional CNN-RNN models by dynamically focusing on different parts of an image while generating a caption that are semantically and syntactically accurate [19]. Wei *et al.* proposed self-attention mechanisms to extract rich image features for captioning of images [20]. The model focusses on sentence level attention and word level attention to provide caption in accordance with the perception of humans. However, using sentence level attention is likely to lead to overfitting and is computationally complex due to double attention mechanisms. The attention weights generated on the image space regions changes continuously and are different at each time step. Hence, the model could not focus on some regions of the image which would reduce the quality of generated captions.

Alahmadi and Hahn [21] attempted to improve the image captioning scores by predicting the sequence of gazing patterns that aligns with human visual perception to generate accurate captions. These predicted sequences are integrated into various image captioning models to significantly improve the image captioning metrics. Further, preserving the order of gazing objects and patterns will be expensive and may lead to lower evaluation scores of the model. The work relies on quality of existing captions and when captions are noisy or unavailable, predicting effective gazing patterns would be compromised. The gazing sequences is a sequential process but to generate quality captions it is necessary to consider different regions of the image in a non-sequential way which can be done by global attention mechanisms.

Semantic ontology in image caption generation was leveraged by Seung-ho Han and Ho-jin Choi [22]. The authors used the concept of dual attention mechanism for detecting objects and identifying attributes to generate contextually relevant captions with semantic understanding. The model is fine-tuned on a data set specific to a domain rather than general dataset like MS COCO and the generic words are replaced with domain specific synonyms. To generalize the model, it requires extensive retraining with new domain specific ontologies. The challenge is that the language decoder using RNN makes it difficult to maintain the context for long sequences and can lead to vanishing or exploding gradient problem. For improved captioning accuracy, the images with multiple objects and complex visual features has to be considered to deploy across broad range of applications.

Image captioning with geometric and position semantics was proposed by Haque *et al.* [23] for avoiding the distorted labels in an image. Position and geometric semantics were used to create descriptions for images and positional encoding methods preserves the spatial details of image in the generated captions. The geometrical properties of images are handled by using parallelised capsule encoding and decoding. The model is tested over flicker8K dataset, and it cannot scale when applied to complex datasets as the computational cost of parallelised capsules will be significantly high. The model lacks evaluation of multiple metrics like SPICE, CIDEr etc., which validates the semantic quality of the captions.

The transformer based architectures uses self-attention mechanisms that models the dependencies between image features and text sequences for generating captions that are contextually rich. The use of transformer architectures for image captioning tasks was discussed by Castro *et al*. [24]. The authors analysed the impact of various hyper-parameter configuration in producing optimal captions with lower computational cost for different transformer architectures. The choice of loss function and optimizer significantly influenced the performance of image captioning models. But it worked well only for specific architectures like vision transformer and data-efficient image transformers. The proposed approach normalizes the captions only to a fixed length whereas dynamic length caption would improve the fluency in image descriptions. Only a subset of hyper-parameters was fine-tuned and better captions could be obtained by tuning additional parameters like model depth, weight decay, size of kernels etc., to the framework.

Li *et al.* [25] proposed a model for enhanced image captioning by combining semantic comprehension and ordering to generate semantic relationship between words. The work incorporates a cross-modal retrieval model, CLIP to retrieve similar sentences based on visual contents of the image. Use of CLIP may introduce biases as it is pre-trained only on specific data sets. The multiple transformer blocks for visual encoding and sematic processing is computationally expensive. For images with new or rare objects, the performance of the model degrades and the few-shot learning techniques can be incorporated for captioning rare or unseen objects. The use of semantic comprehender would filter the diversity in descriptions of images and it may provide repetitive captions which effects the evaluation scores of image captioning models.

An alternative approach is the use of vision transformer based image captioning using grid representations introduced by Fang *et al.* [26]. The vision transformer produces grid features on encoding the image and predicts the semantic concepts of image. The sematic token vocabulary provides concept tokens that are concatenated with the grid representations to give relevant captions. Due to the absence of object detector, the model lacks detailed spatial and attribute information about the image objects and generated captions would be more generic and less specific to the input images. The vision transformer requires extensive training on large pre-trained data sets, which increases the complexity. The model still depends on object based annotations and computationally expensive transformers.

The process of validating the correctness of captions for list of unlabelled images using metamorphic method was presented by Yu *et al.* [27]. The MetaIC framework uses list of unlabelled images and performs object insertions for which the captions are generated. The discrepancies in the captions of the original and object inserted images lead to labelling errors and omission of relevant words in the captions. With relevance to diverse images of the MS COCO dataset, the framework may suffer from errors due to complex scene understanding and produce contextually irrelevant captions. The idea of inserting objects and generating modified images may not work well across diverse models and generating object corpus is also computationally expensive. The generated captions may encounter false positives due to the errors of singular plural form in describing objects.

The testing of image captioning system using reduction based image transformation was proposed by Xie *et al.* [28]. Applying reduction based transformations may cause some objects to be partially cut and it is difficult to give the caption for these ambiguous objects. If the ambiguous captions are deleted, it may reduce the test coverage. The group of test cases are generated based on metamorphic relation using source and follow up images. The useful information in the follow up images will be reduced due to input transformations like cropping, stretching and rotation. There is a dependency on object detection models which may lead to incorrect captions if biased. The model has not explored the error types due to missing objects, hallucinations, misinterpretations which effects the interpretability.

From the above discussion, it is found that the existing models do not generalize well for training large data sets and most models are computationally expensive. Another major challenge is the unavailability of diverse sets of data to build an accurate model. To overcome the above challenges, the proposed STAGRIC model is trained on MS COCO data set which enhances the model's ability to generalize to diverse images. Further, generation of rich sentences would be difficult particularly for long sequences of text due to vanishing gradient problems. Hence, a stacked GAN is designed for training the model using synthetic images along with real images thereby avoiding vanishing and exploding gradient problems. The STAGRIC

architecture is built with Inceptionv3 model and visual attention based GRU model to generate contextually rich captions with reduced computational complexity.

3. Proposed Architecture

Image captioning is the task of describing the content of an image. The goal is to accomplish such a task by building an efficient and accurate model overcoming the challenges cited earlier. Leveraging advanced deep learning techniques, a Stacked GAN and Gated Recurrent Units based Image Caption generator (STAGRIC) is proposed here. Stacked GAN is an advanced version of GAN that can generate images from text where it uses two generators and two discriminators to generate high quality synthetic images. The Inceptionv3 model is a good candidate with pretrained weights that can be used to extract the feature vectors from the images. The decoder model which uses Gated Recurrent Units (GRU) helps in decreasing the computational complexity. The key advantage of the proposed methodology is that with less amount of data an accurate model can be built since the training dataset is augmented with synthetic data. This overcomes the limitation of unavailability of large datasets, otherwise it would have resulted in overfitting or underfitting. The second advantage is that the proposed methodology can produce better models with computationally less intensive training phase.

3.1. STAGRIC Architecture

The system architecture of STAGRIC image caption generator system is depicted in Figure 1. In the image generator module, the text captions from the dataset are given as input to the Text Vectorizer. The text vectorizer produces the text features in numerical form for the model to process. The text features are then encoded and used by the Stacked GAN model to produce synthetic images. These generated synthetic images are mapped to their respective captions. Later, this dataset is added to the original dataset which helps us to have large data in building the caption generation model.

Stacked GAN splits the complex tasks into more manageable subtasks through a sketching and refining process. The elementary shape and colour of the image generated from the text description are sketched by Stage 1 Generator, yielding images which are of low resolution. The Stage 1 results are given to Stage 2 GAN along with text descriptions to generate high resolution images with realistic information. Defects in the Stage 1 results are rectified, and compelling details are added with the process of refinement.

In the caption generator model, the images from the original dataset are pre-processed and its features extracted using Inceptionv3 CNN model. Later, these features form the input to the GRU based Visual Attention Mechanism which generates captions for the image. Inceptionv3 pre-trained model is imported here, and its final layers are modified to get the image features vectors. This model is the encoder model. Then, an attention model is built which takes care of attention weights. Finally, a decoder model is designed using GRU algorithm and the attention weights from the attention model are utilised to generate accurate image captions. The dataset used is MS COCO which is a popular captioning dataset that has over three hundred thousand images annotated with eighty object categories and five captions for each image.

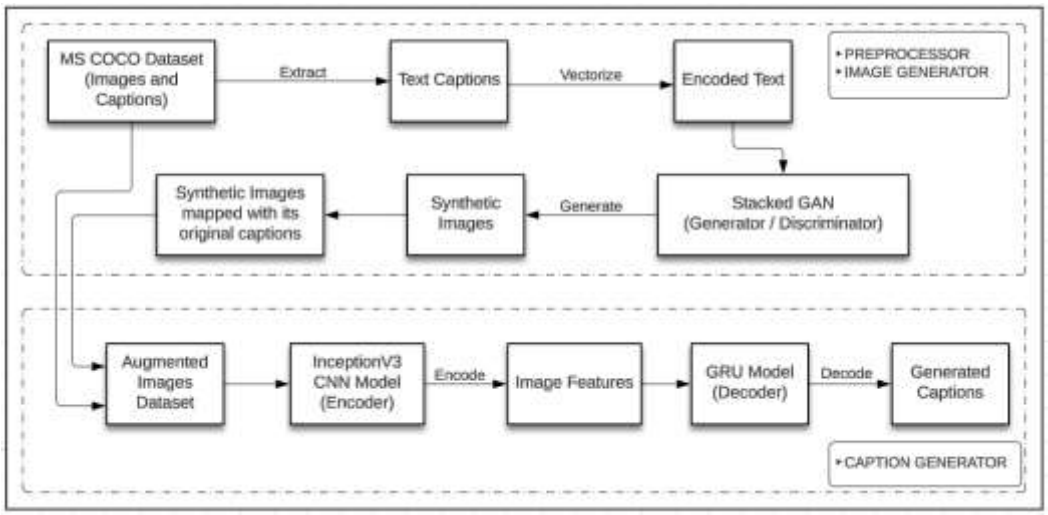

Figure 1. Overview of STAGRIC Architecture

3.2. Image Caption Generator Algorithm built in STAGRIC Architecture

The image caption generator algorithm in the STAGRIC architecture comprises the following sequence of steps:

Step 1: Prepare 20000 entries each having an image and its corresponding text caption from MS COCO dataset.
Step 2: This dataset is loaded into Google Colab, and its contents are extracted.
Step 3: Text captions of the images are extracted, pre-processed, and applied to Text Vectorizer which outputs text features.
Step 4: The dataset is split based on the commonly used standard Karpathy split [29].
Step 5: Using the Inceptionv3 model, the features of the pre-processed images are extracted.
Step 6: Text captions from this dataset are pre-processed and their features are extracted using Text Vectorizer once again.
Step 7: Gated Recurrent Units (GRU) architecture is designed and trained with the training data.
Step 8: Then the model is evaluated with the test data. The generated captions are used to compute the BLEU scores which form the baseline data. The above steps are repeated for consistency.
Step 9: To augment the dataset with synthetic images, captions from the main dataset are selected and synthetic images are generated using trained Stacked GAN.

These synthetic images along with their captions form the augmented dataset against which Steps 1 to 8 are executed for generating image caption and evaluated by computing BLEU scores. In Step 5, Inceptionv3 model is chosen since it is best known pretrained deep model on large datasets. In Step 7, GRU is chosen since it involves two gates and aids in the reduction of computational complexity.

3.3. Core Components of STAGRIC Architecture

The three core components that comprise the proposed STAGRIC architecture are discussed as follows.

3.3.1. Preprocessor and Synthetic Image Generator

The dataset used for the study is MS COCO dataset containing over 20000 images and their corresponding captions. This dataset is imported into Google Colab using Keras libraries. The contents of these files are extracted and stored in two data frames, where one data frame contains the images, and the other one contains the respective captions of the images. All the images in this dataset will be resized to a similar size to maintain the uniformity of the images. The model being used is Inceptionv3 which has already built a pre-processing function, and these images must be applied with that function.

To preprocess the texts, all the text captions in the dataset are made lower-cased, the unwanted characters in the text are replaced and all the texts are tokenized. In each caption, the maximum word is fixed at 50 and the size of the vocabulary is assigned 5000 to avoid computational complexity. The Text Vectorizer algorithm has been used to assign a numerical value for all the words available in the vocabulary. The images are visualised to check whether their sizes are uniform or not. The images are visualised in grayscale to check whether the colour has an influence in the images. Then, the images are visualised along with their captions. Stacked GAN [30] is used to generate synthetic images using the text captions and then added to the dataset so that more data can be processed with less complexity. This methodology is superior to other techniques in generating photo realistic images with less GPU power and hence finds its place in STAGRIC architecture. The original images and synthetic images of a caption are visualised to determine the quality of synthetic images. All these functions are done using the PIL Python library and the pseudo code is shown in Scheme I.

3.3.2. Image Feature Extractor

The goal here is to develop a model which can extract image features or spatial information from the images. Inceptionv3 [31] is a deep network which has higher efficiency in analysing images for generating features and is computationally less expensive. The pseudo code for image feature extraction is illustrated in Scheme II.

The step wise procedure for extraction of image features is given below:

Step 1: Initialise the Inceptionv3 pre-trained model from Keras library.
Step 2: Train the model using Imagenet [32] which is a large dataset consisting of various classes of

images.

Step 3: In the pre-trained model that is imported remove the last layer which is used to classify the images.

Step 4: Add a fully connected Dense layer with a ReLU activation function to extract the feature vector of the images.

Scheme I. Pseudocode for Preprocessor component

```
import tensorflow as tflow
current_image = tflow.io.read_file(path_of_image)
current_image = tflow.io.decode_jpeg(current_image, channels=3)
current_image = tflow.keras.layers.Resizing(299, 299)( current_image)
current_image = tflow.keras.applications.inception_v3.preprocess_input(current_image)

captions_dataset = tflow.data.Dataset.from_tensor_slices(training_captions)
output_seq_max_length = 50
size_of_vocabulary = 5000
tokenizer = tflow.keras.layers.TextVectorization(max_tokens= size_of_vocabulary,
        standardize=standardize, output_sequence_length=output_seq_max_length)
tokenizer.adapt(captions_dataset)
captions_vector = captions_dataset.map(lambda x: tokenizer(x))
```

Scheme II. Pseudocode for Image Feature extractor component

```
import tensorflow as tflow
inception_model = tflow.keras.applications.InceptionV3(include_top=False, weights='imagenet'
input_new = inception_model.input
hidden_layer = inception_model.layers[-1].output
extract_images_features_model = tflow.keras.Model(input_new, hidden_layer)
for each image in dataset: extracted_images_features = extract_images_features_model(image)

class CNN_Encoder(tflow.keras.Model):
self.fc = tflow.keras.layers.Dense(embedding_dimension)
def call(self, x):
x = self.fx(x)
x = tflow.nn.relu(x)
return x
```

3.3.3. Image Caption Generator

Gated Recurrent Units (GRU) based Visual Attention Mechanism [33] is used to produce textual description of the images. By using the Gated Recurrent Units concept, the computational complexity faced in LSTM can be reduced as GRU is a dual-gate system whereas LSTM is a three-gate system. Upon this, the attention mechanism concept would be combined where the higher weights are assigned to the highly suitable words. The key implementation for caption generation is illustrated in Scheme III.

Step 1: Define two classes, one for Attention Mechanism and the other for GRU based Visual attention mechanism.

Step 2: Model the Attention Mechanism class using the Dense layer and activation functions tanh and softmax available in the TensorFlow package.

Step 3: Use the attention weights returned in the previous step in the other class that uses GRU based Visual Attention Mechanism.

Scheme III. Pseudocode for Image Caption Generator component

```
import tensorflow as tflow
class BahdanauAttention(tflow.keras.Model):
self.W1 = tflow.keras.layers.Dense(units)
self.W2 = tflow.keras.layers.Dense(units)
self.V = tflow.keras.layers.Dense(1)

class RNN_Decoder(tflow.keras.Model):
self.embedding = tflow.keras.layers.Embedding(vocabulary_size, embedding_dimension)
self.gru = tflow.keras.layers.GRU(self.units, return_sequences=True, return_state=True,
        recurrent_initializer='glorot_uinform')
self.fc1 = tflow.keras.layers.Dense(self.units)
self.fc2 = tflow.keras.layers.Dense(size_of_vocabulary)
self.attention = BahdanauAttention(self.units)
```

The additive attention mechanism employed here generates an attention score as in equation (1) using both encoder state h_i and decoder hidden s_t. With V_a as weight vector and W_a as weight matrix, the attention score is computed as follows.

$$score(s_t, h_i) = V_a^T \tanh\left(W_a[s_t; h_i]\right) \quad (1)$$

The attention score is passed on to the softmax function to compute the attention weights α as in equation (2).

$$\alpha_{t,i} = softmax(score(s_t, h_i)) \quad (2)$$

Subsequently the context vector c is generated to compute the final output for predicting the word as in equation (3).

$$c_t = \sum_{i=1}^{T} \alpha_{t,i}\, h_i \quad (3)$$

To summarise, in GRU based Visual attention mechanism class, Embedding, GRU and Dense Layer from the TensorFlow package have been used and the weights from the attention mechanism class have been utilised to build the complete model which can generate the captions for images.

4. Results and Discussion

4.1. Experimental Setup

The STAGRIC architecture is evaluated using the MS COCO dataset, a publicly available dataset useful for researchers in various fields dealing with images. This dataset is composed of over 300,000 general images spread across multiple categories. All these images have five text captions each. The software libraries used for conducting the experiment runs in Google Colab [34] which is a hosted Jupyter notebook service on the cloud. The runtime environment was set to use T4 GPU which showed significant performance improvement compared to CPU and TPU since the libraries were optimised for GPU. Keras and TensorFlow [35-36] are Python libraries that are used hand in hand for conducting most of the deep learning experiments. While TensorFlow does the heavy lifting on the compute side to optimally execute various algorithms, Keras is the front end to interact with TensorFlow.

The source code of the experiment is available in https://github.com/navincj/image_captioning for the benefit of other researchers. Table 1 summarizes all the design parameters that was applied to conduct the experiments. With the size of the batch set to 64, the training phase was run for 100 epochs for convergence. Adam was used as the optimizer algorithm and for assigning the initial weights to the neural network parameter, Xavier/Glorot technique was adopted. The learning rate of 0.00005 is used in Stacked GAN which is used for generating synthetic images.

To validate the hypothesis, the datasets are filtered by specific set of categories to retrieve a manageable volume of images. With this strategy, the dataset containing over 20000 entries each having an image, and its corresponding caption is prepared. The existing ground truth captions are used to generate synthetic images to augment the dataset. Adopting STAGRIC architecture, models are trained to produce improved text captions based on the image given as input. The results are tabulated and compared with other models. Additionally, ablation study conducted for the proposed model has resulted in 25% improvement in BLEU metrics highlighting the significance of synthetic datasets.

Table 1. STAGRIC Experimental Settings

Experimental Components	Parameters and Values
Dataset	MS COCO
Framework	Google Colab, Tensorflow, Keras
Training Epochs	100
Size of the Batch	64
Learning Rate	0.0001
Optimizer	Adam
Weight Initialization	Xavier/Glorot
Stacked GAN Training Steps	5 (Used in Generator and Discriminator update)
Stacked GAN Learning Rate	0.00005

Figure 2 lists a few random samples of synthetic images generated from caption using Stacked GAN. This key feature of STAGRIC architecture helps in handling domains with limited data. The dual stage in Stacked GAN plays a key role in generating realistic images capturing the key aspects of the input captions

and is quite evident in the samples. Figure 3 lists a few samples of the generated captions from the trained models of the STAGRIC architecture. The augmented dataset used in the architecture generates superior model and all the regions of interest in the image are vividly captured in the generated captions.


A cat that is laying down on a bag.


A living room, with the TV on in the background.


A tall clock tower with a sky background.


A bike parked on the road.

Figure 2. Synthetic Images generated from captions using Stacked GAN


A person with a pink umbrella on the sidewalk.


Two horses on a buggy in a crowded waterside market.


A man is walking down a skateboard.


A person with a backpack skiing in the snow.

Figure 3. Captions generated from images using STAGRIC

4.2. Interpretation of Results

Table 2 summarises the image caption evaluation metrics [37] for various datasets used in STAGRIC along with metrics generated from other popular architectures for comparison. The commonly used standard Karpathy split was adopted on the datasets to split for training and testing. It is evident from the results that STAGRIC outperforms the other models considered for the study. The evaluation was further extended by subjecting the STAGRIC model to generate captions by using original images exclusively as well as original images augmented with synthetic images. The presence of synthetic images boosts the caption metrics consistently across all the parameters. The high BLEU-1 score of 75.8 computed in the proposed STAGRIC architecture clearly demonstrates significant improvement compared to the models that used only original images. Synthetic images generated using Stacked GAN are computationally less intensive as well and is of high quality. The augmented dataset used in the architecture yields superior captions which is reflected consistently in all the metrics.

Table 2. Image Caption Metrics for various Models

Models	Dataset	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGH-L
ICPS [23]	Original Images only	61.4	38.7	23.5	24.0	24.5	43.7
COS-Net [25]	Original Images only	72.7	52.8	32.6	30.3	24.6	53.9
STAGRIC	Original Images only	62.3	41.8	26.7	17.0	19.7	43.2
STAGRIC	Original and Synthetic Images	**75.8**	**55.1**	**39.8**	**37.4**	**29.6**	**55.2**

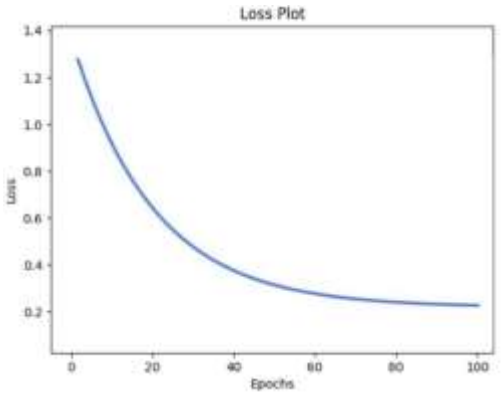

Figure 4. STAGRIC with original images

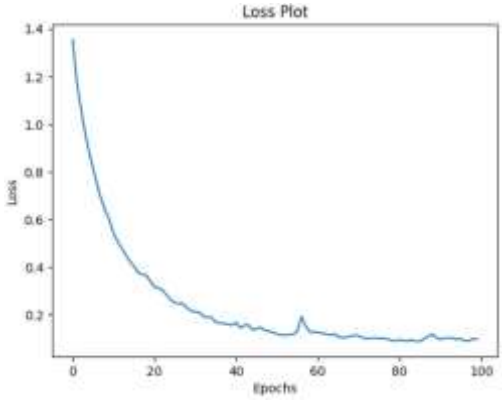

Figure 5. STAGRIC with original and synthetic images

Loss plot in Deep Learning models signifies the efficacy of the training phase. Figure 4 depicts the loss plot against 100 epochs in X-axis for the model trained with only the original dataset. This is captured to compare with the STAGRIC architecture where both original and synthetic images are used. Figure 5

captures the loss plot when the model is trained with STAGRIC architecture using both original and synthetic images. It is noticed that loss settles around 10% in the first plot whereas it goes further less than 5% in STAGRIC architecture trained with both original and synthetic images. Comparing the results, a marked improvement in terms of reduction in loss is noticed highlighting the effectiveness of synthetic images to generate superior image caption generator models. All the experimental results against random sample sets corroborate the same. The study will find enormous interest among medical fraternity and its allied domains where high accuracy is expected and are often constrained by limited training data.

5. Conclusion and Future Work

Image Caption generation has kindled the interest of researchers in Deep Learning and finds numerous applications in real world systems that includes biomedical imaging, surveillance systems, assistance for visually impaired people etc. Often, the unavailability of a large volume of datasets causes challenges to build a more generalised accurate model. It is observed that the model built using datasets containing both original and synthetic images outperform significantly compared to having just the original datasets. The novel Image Caption Generator system uses stacked GAN to generate synthetic images and GRU based Visual Attention Mechanism (STAGRIC) is an effective and computationally efficient approach to generate descriptive text captions for any image. The hypothesis has been successfully validated with improved BLEU scores of the generated image captions using MS COCO dataset. This work successfully demonstrates the efficacy of the novel STAGRIC architecture to build an efficient image caption generator with limited data. The outcome of this research will immensely help in the domains that usually have limited data.

As future work, the evaluation of different algorithms to generate synthetic images to boost the BLEU scores may be considered. Another area to delve will be to include semantic ontology in caption generation. This would be a logical extension of the work to highlight the role played by semantics to generate superior captions. Leveraging the deeper understanding of images to study the detection of deep fakes could be an interesting area to explore.

CRediT Author Contribution Statement

Navin Chandar Jacob: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Resources, Software, Validation, Visualization, Writing—Original Draft and Writing—Review & Editing; Kavitha Ganesh: Project administration, Supervision and Validation; Aakash Sethuraman: Investigation, Software and Resources.

References

[1] Hyunseok Seo, Masoud Badiei Khuzani, Varun Vasudevan, Charles Huang, Hongyi Ren *et al.*, "Machine Learning Techniques for Biomedical Image Segmentation: An Overview of Technical Aspects and Introduction to State-of-Art Applications", *Medical physics*, Online ISSN: 2473-4209, pp. 148-167, Vol. 47, No. 5, 17th May 2020, Published by The International Journal of Medical Physics Research and Practice, DOI: 10.1002/mp.13649, Available: https://aapm.onlinelibrary.wiley.com/doi/abs/10.1002/mp.13649.

[2] Shruti Agarwal, Hany Farid, Tarek El-Gaaly and Ser-Nam Lim, "Detecting Deep-Fake Videos From Appearance And Behavior", in *Proceedings of the IEEE International Workshop on Information Forensics And Security (WIFS)*, 06-11 December 2020, New York, USA, Online ISBN: 978-605-01-1275-7, Print ISBN: 978-1-7281-2394-3, pp. 1-6, Published by IEEE, DOI: 10.1109/WIFS49906.2020.9360904, Available: https://ieeexplore.ieee.org/document/9360904.

[3] Alexey Stulov, Andrey Tikhonov and Irina Snitko, "Fundamentals of Artificial Intelligence in Power Transformers Smart Design", in *Proceedings of the International Ural Conference on Electrical Power Engineering (UralCon)*, 22-24 September 2020, Chelyabinsk, Russia, E-ISBN: 978-1-7281-6209-6, USB ISBN: 978-1-7281-6208-9, Print ISBN: 978-1-7281-6210-2, pp. 34-38, Published by IEEE, DOI: 10.1109/UralCon49858.2020.9216245, Available: https://ieeexplore.ieee.org/document/9216245.

[4] Janmanchi Harika, Palavadi Baleeshwar, Kummari Navya and Hariharan Shanmugasundaram, "A Review on Artificial Intelligence with Deep Human Reasoning", in *Proceedings of the International Conference on Applied Artificial Intelligence and Computing (ICAAIC)*, 9-11 May 2022, Salem, India, E-ISBN: 978-1-6654-9710-7, DVD ISBN: 978-1-6654-9709-1, Print ISBN: 978-1-6654-9711-4, pp. 81-84, Published by IEEE, DOI: 10.1109/ICAAIC53929.2022.9793310, Available: https://ieeexplore.ieee.org/document/9793310.

[5] Chen Chen, Shuai Mu, Wanpeng Xiao, Zexiong Ye, Liesi Wu and Qi Ju, "Improving Image Captioning with Conditional Generative Adversarial Nets", in *Proceedings of the AAAI conference on artificial intelligence*, Vol. 33, No. 01, 27 January – 1 February 2019, Honolulu, Hawaii, USA, Online ISSN: 2374-3468, Print ISSN: 2159-5399, pp. 8142-8150, Published by AAAI Press, DOI: 10.1609/aaai.v33i01.33018142, Available: https://ojs.aaai.org/index.php/AAAI/issue/view/246.

[6] Mohamed Omri, Sayed Abdel-Khalek, Eied M. Khalil, Jamel Bouslimi and Gyanendra Prasad Joshi, "Modeling of Hyperparameter Tuned Deep Learning Model for Automated Image Captioning", *Mathematics*, Online ISSN: 2227-7390, pp. 288, Vol. 10, No. 3, 2022, Published by MDPI, DOI:10.3390/math10030288, Available: https://www.mdpi.com/2227-7390/10/3/288.

[7] Kalpana Deorukhkar, Satish Y. Ket, "A detailed review on Prevailing Image Captioning Methods using Deep Learning Techniques", *Multimedia Tools and Applications*, Online ISSN: 1380-7501, pp. 1313-1336, Vol. 81, No.1, January 2022, Published by Springer, DOI:10.1007/s11042-021-11293-1, Available: https://dl.acm.org/toc/mtaa/2022/81/1.

[8] Wei Zhao, Wei Xu, Min Yang, Jianbo Ye, Zhou Zhao *et al.*, "Dual Learning for Cross-Domain Image Captioning", in *Proceedings of the ACM on Conference on Information and Knowledge Management*, 6-10 November 2017, Singapore, Singapore, ISBN: 978-1-4503-4918-5, pp. 29-38, Published by the Association for Computing Machinery, DOI: 10.1145/3132847.3132920, Available: https://dl.acm.org/doi/10.1145/3132847.3132920.

[9] Mohamed Ragab, Zhenghua Chen, Min Wu, Chee-Keong Kwoh, Ruqiang Yan, *et al.*, "Attention-based Sequence to Sequence Model for Machine Remaining Useful Life Prediction", *Neurocomputing*, Online ISSN: 1872-8286, pp. 58-68, Vol. 466, 27th November 2021, Published by Elsevier, DOI: 10.1016/j.neucom.2021.09.022, Available: https://www.sciencedirect.com/science/article/pii/S0925231221013801.

[10] J Sudhakar, Viswesh V. Iyer and Sree T. Sharmila, "Image Caption Generation Using Deep Neural Networks", in *Proceedings of the International Conference for Advancement in Technology (ICONAT)*, 21-22 January 2022, Goa, India, E-ISBN: 978-1-6654-2577-3, Print ISBN: 978-1-6654-2578-0, pp. 1-3, Published by IEEE, DOI: 10.1109/ICONAT53423.2022.9726074, Available: https://ieeexplore.ieee.org/document/9726074.

[11] Akash Kothare, Shridhara Chaube, Yash Moharir, Gaurav Bajodia and Snehlata Dongre, "SynGen: Synthetic Data Generation", in *Proceedings of the International Conference on Computational Intelligence and Computing Applications (ICCICA)*, 26-27 November 2021, Nagpur, India, E-ISBN: 978-1-6654-2040-2, Print ISBN: 978-1-6654-2041-9, pp. 1-4, Published by IEEE, DOI: 10.1109/ICCICA52458.2021.9697232, Available: https://ieeexplore.ieee.org/document/9697232.

[12] Renana Peres, Martin Schreier, David Schweidel and Alina Sorescu, "On ChatGPT and Beyond: How Generative Artificial Intelligence May Affect Research, Teaching, And Practice", *International Journal of Research in Marketing*, Online ISSN: 1873-8001, pp. 269-275, Vol. 40, No. 2, 9th June 2023, Published by Elsevier, DOI: 10.1016/j.ijresmar.2023.03.001, Available: https://www.sciencedirect.com/science/article/pii/S0167811623000162.

[13] Gabriel Hermosilla, Diego-Ignacio Henríquez Tapia, Hector Allende-Cid, Gonzalo Farías Castro and Esteban Vera, "Thermal Face Generation using Stylegan", *IEEE Access*, E-ISSN: 2169-3536, pp. 80511-80523, Vol. 9, 02 June 2021, Published by IEEE, DOI: 10.1109/ACCESS.2021.3085423, Available: https://ieeexplore.ieee.org/document/9445031.

[14] Bruno Vaz, Álvaro Figueira, "GANs in the Panorama of Synthetic Data Generation Methods", *ACM Transactions on Multimedia Computing, Communications and Applications*, ISSN: 1551-6857, E-ISSN:1551-6865, pp. 1-28, Vol. 21, No. 1, 14 December 2024, Published by the Association for Computing Machinery, DOI: 10.1145/3657294, Available: https://dl.acm.org/doi/full/10.1145/3657294.

[15] Aisha Zulfiqar, Sher Muhammad Daudpota, Ali Shariq Imran, Zenun Kastrati, Mohib Ullah *et al.*, "Synthetic Image Generation Using Deep Learning: A Systematic Literature Review", *Computational Intelligence*, Online ISSN: 1467-8640, Print ISSN: 0824-7935, Vol. 20, No. 5, 21 October 2024, Published by Wiley Periodicals, DOI: 10.1111/coin.70002, Available: https://onlinelibrary.wiley.com/doi/10.1111/coin.70002.

[16] Jafar A. Alzubi, Rachna Jain, Preeti Nagrath, Suresh Satapathy, Soham Taneja *et al.*, "Deep image captioning using an ensemble of CNN and LSTM based deep neural networks", *Journal of Intelligent & Fuzzy Systems: Applications in Engineering and Technology*, ISSN: 1064-1246, pp. 5761-5769, Vol. 40, No. 4, 1 January 2024, Published by the Association for Computing Machinery, DOI: 10.3233/JIFS-189415, Available: https://dl.acm.org/doi/10.3233/JIFS-189415.

[17] Md Zakir Hossain, Ferdous Sohel, Mohd Fairuz Shiratuddin, Hamid Laga and Mohammed Bennamoun "Attention-Based Image Captioning using DenseNet Features", in *Proceedings of the 26th International Conference on Neural Information Processing, ICONIP 2019, Proceedings, Part V 26*, 12-15 December 2019, Sydney, NSW, Australia, Print ISBN: 978-3-030-36801-2, Online ISBN: 978-3-030-36802-9, pp. 109-117, Published by Springer International, DOI: 10.1007/978-3-030-36802-9_13, Available: https://link.springer.com/chapter/10.1007/978-3-030-36802-9_13.

[18] Akash Verma, Harshit Saxena, Mugdha Jaiswal and Poonam Tanwar, "Intelligence Embedded Image Caption Generator using LSTM based RNN Model", in *Proceedings of the 6th International Conference on Communication and Electronics Systems (ICCES)*, 8-10 July 2021, Coimbatore, India, E-ISBN: 978-1-6654-3587-1, Print ISBN: 978-1-6654-

1182-0, pp. 963-967, Published by IEEE, DOI: 10.1109/ICCES51350.2021.9489253, Available: https://ieeexplore.ieee.org/document/9489253.

[19] Uday Kulkarni, Kushagra Tomar, Mayuri Kalmat, Rakshita Bandi, Pranav Jadhav *et al.*, "Attention based Image Caption Generation (ABICG) using Encoder-Decoder Architecture", in *2023 5th International Conference on Smart Systems and Inventive Technology (ICSSIT)*, 23-25 January 2023, Tirunelveli, India, Electronic ISBN: 978-1-6654-7467-2, Electronic ISSN: 2832-3017, Published by IEEE, DOI: 10.1109/ICSSIT55814.2023.10061040, Available: https://ieeexplore.ieee.org/document/10061040.

[20] Haiyang Wei, Zhixin Li, Canlong Zhang, Tao Zhou and Yu Quan, "Image Captioning Based on Sentence-Level and Word-Level Attention", in *Proceedings of the International Joint Conference on Neural Networks (IJCNN)*, 14-19 July 2019, Budapest, Hungary, E-ISBN: 978-1-7281-1985-4, Print ISBN: 978-1-7281-1986-1, E-ISSN: 2161-4407, Print ISSN: 2161-4393, pp. 1-8, Published by IEEE, DOI: 10.1109/IJCNN.2019.8852118, Available: https://ieeexplore.ieee.org/document/8852118.

[21] Rehab Alahmadi and James Hahn, "Improve Image Captioning by Estimating the Gazing Patterns from the Caption", in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 3-8 January 2022, Waikoloa, USA, E-ISBN: 978-1-6654-0915-5, Print ISBN: 978-1-6654-0916-2, E-ISSN: 2642-9381, Print ISSN: 2472-6737, pp. 2453-2462, Published by IEEE, DOI: 10.1109/WACV51458.2022.00251, Available: https://ieeexplore.ieee.org/document/9707074.

[22] Seung-Ho Han and Ho-Jin Choi, "Domain-Specific Image Caption Generator with Semantic Ontology", in *Proceedings of the IEEE International Conference on Big Data and Smart Computing (BigComp)*, 19-22 February 2020, Busan, South Korea, E-ISBN: 978-1-7281-6034-4, Print ISBN: 978-1-7281-6035-1, E-ISSN: 2375-9356, Print ISSN: 2375-933X, pp. 526-530, Published by IEEE, DOI: 10.1109/BigComp48618.2020.00-12, Available: https://ieeexplore.ieee.org/document/9070680.

[23] Anwar Ul Haque, Sayeed Ghani and Muhammad Saeed, "Image Captioning with Positional and Geometrical Semantics", *IEEE Access*, E-ISSN: 2169-3536, pp. 160917-160925, Vol. 9, 30th November 2021, Published by IEEE, DOI: 10.1109/ACCESS.2021.3131343, Available: https://ieeexplore.ieee.org/document/9627971.

[24] Roberto Castro, Israel Pineda, Wansu Lim, Manuel Eugenio Morocho-Cayamcela, "Deep Learning Approaches Based on Transformer Architectures for Image Captioning Tasks", *IEEE Access*, E-ISSN: 2169-3536, pp. 33679-33694, Vol. 10, 22nd March 2022, Published by IEEE, DOI: 10.1109/ACCESS.2022.3161428, Available: https://ieeexplore.ieee.org/document/9739703.

[25] Li Yehao, Yingwei Pan, Ting Yao and Tao Mei, "Comprehending and Ordering Semantics for Image Captioning", in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 18-24 June 2022, New Orleans, USA, E-ISBN: 978-1-6654-6946-3, Print ISBN: 978-1-6654-6947-0, E-ISSN: 2575-7075 , Print ISSN: 1063-6919, pp. 17969-17978, Published by IEEE, DOI: 10.1109/CVPR52688.2022.01746, Available: https://ieeexplore.ieee.org/document/9879735.

[26] Zhiyuan Fang, Jianfeng Wang, Xiaowei Hu, Lin Liang, Zhe Gan *et al.*, "Injecting Semantic Concepts into End-to-End Image Captioning", in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 18-24 June 2022, New Orleans, USA, E-ISBN: 978-1-6654-6946-3, Print ISBN: 978-1-6654-6947-0, E-ISSN: 2575-7075, Print ISSN: 1063-6919, pp. 17988-17998, Published by IEEE, DOI: 10.1109/CVPR52688.2022.01748, Available: https://ieeexplore.ieee.org/document/9879403.

[27] Boxi Yu, Zhiqing Zhong, Xinran Qin, Jiayi Yao, Yuancheng Wang *et al.*, "Automated testing of image captioning systems", in *ISSTA 2022: Proceedings of the 31st ACM SIGSOFT International Symposium on Software Testing and Analysis*, 18-22 July 2022, Virtual, South Korea, ISBN: 978-1-4503-9379-9, pp. 467–479, Published by the Association for Computing Machinery, DOI: 10.1145/3533767.3534389, Available: https://dl.acm.org/doi/10.1145/3533767.3534389.

[28] Xiaoyuan Xie, Xingpeng Li and Songqiang Chen, "Metamorphic Testing of Image Captioning Systems via Image-Level Reduction", *IEEE Transactions on Software Engineering*, Print ISSN: 0098-5589, pp. 2962-2982, Vol. 50, No. 11, November 2024, Published by IEEE, DOI: 10.1109/TSE.2024.3463747, Available: https://www.computer.org/csdl/journal/ts/2024/11/10684067/20okCAw7mg0.

[29] Andrej Karpathy and Li Fei-Fei, "Deep Visual-Semantic Alignments for Generating Image Descriptions", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Print ISSN: 0162-8828, E-ISSN: 1939-3539, pp. 664-676, Vol. 39, No. 4, 05 August 2016, Published by IEEE, DOI: 10.1109/TPAMI.2016.2598339, Available: https://ieeexplore.ieee.org/document/7534740.

[30] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang *et al.*, "Stackgan++: Realistic Image Synthesis with Stacked Generative Adversarial Networks", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Print ISSN: 0162-8828, E-ISSN: 1939-3539, pp. 1947-1962, Vol. 41, No. 8, 16th July 2018, Published by IEEE, DOI: 10.1109/TPAMI.2018.2856256, Available: https://ieeexplore.ieee.org/document/8411144.

[31] Niharika Abhange, Swarad Gat and Shilpa Paygude, "COVID-19 Detection using Convolutional Neural Networks and InceptionV3", in *Proceedings of the 2nd Global Conference for Advancement in Technology (GCAT)*, 1-3 October 2021,

Bangalore, India, E-ISBN: 978-1-6654-1836-2, Print ISBN: 978-1-6654-3070-8, pp. 1-5, Published by IEEE, DOI: 10.1109/GCAT52182.2021.9587744, Available: https://ieeexplore.ieee.org/document/9587744.

[32] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li *et al.*, "Imagenet: A Large-Scale Hierarchical Image Database", in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 20-25 June 2009, Miami, USA, Print ISBN: 978-1-4244-3992-8, pp. 248-255, Published by IEEE, DOI: 10.1109/CVPR.2009.5206848, Available: https://ieeexplore.ieee.org/document/5206848.

[33] Sarthak Singh Rawat, Kartikeyan Singh Rawat and Rahul Nijhawan, "A Novel Convolutional Neural Network-Gated Recurrent Unit Approach for Image Captioning", in *Proceedings of the Third International Conference on Smart Systems and Inventive Technology (ICSSIT)*, 20-22 August 2020, Tirunelveli, India, E-ISBN: 978-1-7281-5821-1, Print ISBN: 978-1-7281-5822-8, pp. 704-708, Published by IEEE, DOI: 10.1109/ICSSIT48917.2020.9214109, Available: https://ieeexplore.ieee.org/document/9214109.

[34] Ekaba Bisong, Google Colaboratory, in Building Machine Learning and Deep Learning Models on Google Cloud Platform – A Comprehensive Guide for Beginners, 1st ed. Berkeley, USA: Apress, 28th September 2019, ch. 7, pp. 59-64, Print ISBN: 978-1-4842-4469-2, Online ISBN: 978-1-4842-4470-8, DOI: 10.1007/978-1-4842-4470-8_7, Available: https://link.springer.com/chapter/10.1007/978-1-4842-4470-8_7.

[35] Nikhil Ketkar, "Introduction to Keras", in Deep Learning with Python, 1st ed. Berkeley, USA: Apress, 05th October 2017, ch. 7, pp. 97-111, Print ISBN: 978-1-4842-2765-7, Online ISBN: 978-1-4842-2766-4, DOI: 10.1007/978-1-4842-2766-4_7, Available: https://link.springer.com/chapter/10.1007/978-1-4842-2766-4_7.

[36] Bo Pang, Erik Nijkamp and Ying Nian Wu, "Deep Learning with Tensorflow: A Review", *Journal of Educational and Behavioral Statistics*, ISSN: 1076-9986, Online ISSN: 1935-1054, pp. 227-248, Vol. 45, No. 2, 10th September 2019, Published by Sage Publishing, DOI: 10.3102/10769986198727, Available: https://journals.sagepub.com/doi/10.3102/1076998619872761.

[37] Kishore Papineni, Salim Roukos, Todd Ward and Wei-Jing Zhu, "BLEU: A Method for Automatic Evaluation of Machine Translation", in *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, 7-12 July 2002, Philadelphia, USA, pp. 311-318, Published by the Association for Computing Machinery (ACM), DOI: 10.3115/1073083.1073135, Available: https://dl.acm.org/doi/10.3115/1073083.1073135.