International Journal of Computer Engineering in Research Trends
ISSN: 2349-7084/ https://doi.org/10.22362/ijcert.v11i1.919
Volume 11, Issue 1, January 2024

Research Paper

A Deep Learning Model for Automatic Image Captioning using GRU and Attention Mechanism

Sushma Jaiswal[1*], Harikumar Pallthadka[2], Rajesh P. Chinchewadi[3], Tarun Jaiswal[4]

[1]*Guru Ghasidas Central University, Bilaspur (C.G.) and Post-Doctoral Research Fellow, Manipur International University, Imphal, Manipur, jaiswal1302@gmail.com, https://orcid.org/0000-0002-6253-7327*
[2]*Manipur International University, Imphal, Manipur, vc@miu.edu.in, https://orcid.org/0000-0002-0705-9035.*
[3]*Manipur International University, Imphal, Manipur, rajesh.cto@miu.edu.in.*
[4] *National Institute of Technology, Raipur, tjaiswal_1207@yahoo.com, https://orcid.org/0000-0003-3963-4548*

**Sushma Jaiswal: jaiswal1302@gmail.com*

Received: 11/10/2023, Revised: 12/11/2023, Accepted: 26/12/2023 Published: 03/01/2024

Abstract: - *In computer vision and natural language processing, automatic image captioning is an important task that aims to produce accurate and meaningful image captions. For automatic image captioning, we provide a novel method in this paper that combines a deep learning model based on Gated Recurrent Units (GRUs) with an attention mechanism. During the caption generation process, the model can concentrate on pertinent areas of the image by using the attention mechanism, which dynamically weighs the image attributes. This facilitates better matching of the generated captions' related words with the image features. Recurrent neural networks of the GRU kind are used to simulate the sequential structure of natural language and accurately represent word relationships in the output captions. The network learns to produce logical and contextually appropriate descriptions for different kinds of images by being trained on a broad collection of photos and captions. We show that the suggested model is capable of producing high-quality captions by evaluating it using common metrics as BLEU, METEOR, ROUGE, and CIDEr. The results of our experiments demonstrate that our method performs better than baseline methods, demonstrating the benefits of using GRU and an attention mechanism in the image captioning process. The approach is extremely relevant in real-world applications like image interpretation, accessibility, and content recommendation since the generated captions are not only correct but also express a deeper knowledge of the visual content in the images.*

Keywords- *GRU, CNN, LSTM, Deep learning, image caption*.

---

1. Introduction

Automatic image captioning, a difficult computer vision and natural language processing challenge, generates cohesive and meaningful captions. It is important in accessibility technology, content recommendation, and image understanding. Recent deep learning advances have led to complex models for this task, with attention mechanisms and recurrent neural networks enhancing captioning accuracy and con-textual relevance. In this study, we propose an innovative approach that leverages both an attention mechanism and a Gated Recurrent Unit (GRU) based deep learning model for auto-matic image captioning. The attention mechanism enables the model to dynamically focus on salient regions of the image, allowing for improved alignment between image features and corresponding words in the generated captions. This dynamic attention fosters a better understanding of the image's content and context during the captioning process.

On the other hand, the GRU, a variant of the traditional recurrent neural network, excels in capturing sequential dependencies in data. By integrating GRU into our model, we aim to address the sequential and contextual nature of natural language, facilitating the generation of captions that are not only linguistically coherent but also contextually accurate. The GRU component enhances the model's ability to learn long-term dependencies and produce more contextually meaningful descriptions.

The primary objective of this study is to explore the synergistic potential of integrating the attention mechanism and GRU-based model. We anticipate that this fusion will significantly enhance the model's performance in generating accurate and con-textually relevant captions for a diverse range of images. The proposed model's effectiveness is evaluated through comprehensive

experiments and benchmarked against baseline models using standard evaluation metrics.
The primary contribution or effort of this paper is as follows:

- With an attention mechanism, the model can dynamically focus on certain areas of the input image while creating captions. This dynamic attention technique aligns image attributes and caption words, allowing the model to collect fine-grained details and important context. Thus, image descriptions are more accurate and contextual.
- GRU, a recurrent neural network variation, captures long-term word relationships and handles natural language's sequential character. GRU efficiently models caption language structure, generating more coherent sentences. GRU learns sequential patterns in captions to generate grammatically correct and contextually acceptable captions.
- Success depends on the attention mechanism's synergy with the GRU-based model. The attention technique improves image feature relevance at each caption generation stage, while the GRU simulates linguistic sequential relationships. The model generates captions that accurately reflect the image while preserving a coherent and logical sentence structure via this fusion.

Attention mechanisms and GRUs improve automatic image captioning, as shown by the proposed model's experimental performance. The generated captions increase accuracy, relevance, and understanding of image visual content, demonstrating the model's potential for various real-world applications.

The rest of this paper is structured as follows: Section 2 provides an overview of related work in image captioning, highlighting existing approaches and their contributions. Section 3 presents the methodology, detailing the proposed model architecture and components. Section 4 describes the experimental setup and evaluation metrics, followed by a discussion of the results in Section 5. Finally, Section 6 concludes the study, summarizing key findings and outlining future directions in automatic image captioning.

2. Related Work

Iandola et al. [1] introduced SqueezeNet, a deep learning architecture that achieves AlexNet-level accuracy while lowering model size and parameters. Through unique architecture, SqueezeNet achieves outstanding performance with fewer parameters than typical deep neural networks. This innovative approach solves model size and resource-constrained deployment problems, making it a major advance in deep learning and neural network architecture.

This architecture emphasises realistic CNN design standards, achieving high performance while lowering computational complexity and memory footprint. The study covers the essential balance between computational efficiency and model correctness, providing insights and guidelines for CNN design. The discoveries improve deep learning and neural network topologies for real-world applications [2].

To optimise model size and accuracy, inverted residuals and linear bottlenecks are prioritised. MobileNetV2 is suited for resource-constrained applications due to its superior computational efficiency and accuracy. This research advances efficient CNN architecture design with its findings [3].

Tan et al. [4] offer a novel neural architecture search method. MnasNet optimises mobile CNN architectures via platform-aware design. MnasNet finds efficient and effective mobile network designs by addressing platform constraints during architectural search. This addition greatly impacts mobile-friendly CNN designs.

Scalable image recognition by Zoph et al. [5] is novel. The research optimizes and automates image recognition neural network creation by learning transferable structures through neural architecture search. The proposed methodology advances scalable and adaptable convolutional neural network (CNN) architectures by discovering architectures that may be easily transferred and changed to varied image recognition tasks.

Novel neural network architecture by Chen et al. [6] Dual path networks optimize feature propagation and information flow. This improves feature reuse and deep network training. Dual pathways increase neural network accuracy and efficiency, making them a major achievement.

Xie et al. [7] offer a novel deep neural network performance improvement method. Aggregated residual transformations use residual connections to improve model correctness. This new aggregation technique uses feature reuse to improve deep neural network topologies. Deep learning for computer vision advances significantly in the paper.

The authors [8] introduced Wide Residual Networks (WRNs) to improve performance by improving network layers. Model capacity increases with widening, improving learning and generalization. Wider networks boost deep networks' representational strength and enable state-of-the-art performance in diverse tasks, according to the study. WRNs promote deep learning.

Wide Residual Networks (WRNs) increase model capacity and learning by spreading network layers. Widening the network improves performance, resulting in state-of-the-art results in numerous tasks, according to the study. WRNs promote deep neural network design, demonstrating the relevance of larger topologies for better representation and results [9].

Diverse network paths and actions improve model robustness and accuracy, according to the authors. The study emphasizes structural variety in very deep networks, a revolutionary approach to network building. PolyNet advances network architectures by showing how structural variety improves computer vision performance [10].

The authors [11] introduce Inception-v4, an upgraded Inception architecture, and investigate residual connections

in Inception networks to create Inception-ResNet. They discovered how Inception modules and residual connections work together to improve training efficiency and network performance. The findings help explain network design and residual connection integration in deep learning architectures.

The authors [12] proposed residual learning, which uses fast connections to train extraordinarily deep neural networks. Networks with hundreds of layers could be built while minimizing degradation. ResNet changed computer vision by achieving world-class image recognition performance. Innovating leftover blocks and shortcut connections shaped deep neural network design and optimization research.

In 2014, Simonyan and Zisserman [13] introduced the Very Deep Convolutional Network, a revolutionary deep learning architecture. A rigorous review of network depth's impact on image recognition performance shows its importance in improving accuracy. The VGG architecture's many layers and consistent convolutional operations performed well on large-scale image recognition tasks. Modular and scalable, it set the standard for deep learning models. The paper's findings affected deep convolutional neural network research, advancing computer vision.

This study presented DenseNet, a new deep learning architecture. The authors suggest a densely connected network with feed-forward connections between layers. This dense connectivity reduces parameters and improves network feature reuse and propagation. DenseNet reduced vanishing gradients, enhanced parameter efficiency, and enabled feature reuse. The research sheds light on network topologies and advances deep learning and computer vision by demonstrating the benefits of highly connected networks [14]. DenseNet, a deep learning architecture by Huang et al. (2017), has dense layer connections. Information flow, feature reuse, and gradient vanishing are improved by this architecture. DenseNet performed well in computer vision challenges, demonstrating its influence on modern convolutional neural network design and its capacity to improve parameter efficiency and accuracy [14].

The authors [15] presented a network with numerous parallel convolutional paths to efficiently collect features at different scales. Inception's improved performance and efficiency set a benchmark for deep learning models and inspired neural network design. Multi-scale feature extraction was greatly impacted by this work in computer vision and deep learning.

Batch Normalization, developed by Ioffe and Szegedy (2015), reduces internal covariate shift and speeds up deep neural network training. This technique stabilizes and accelerates training by normalizing intermediate activations, improving convergence and allowing higher learning rates. Batch Normalization improves training efficiency and is now typical in deep neural network training [16].

AlexNet, a ground-breaking new architectural framework, was presented by the author [17]. On the ImageNet dataset, this deep convolutional neural network was able to accomplish a substantial advance in image classification. Deep learning, in particular convolutional neural networks, was shown to have significant potential by the authors in the context of tackling difficult computer vision problems. The victory of AlexNet over the other competing neural networks in the ImageNet challenge was a watershed moment for the industry, since it laid the groundwork for the deep learning revolution in computer vision.

Liu et al. [18] introduces a novel approach for image caption generation utilizing a dual attention mechanism. The proposed model employs attention mechanisms to dynamically highlight relevant image regions and words during the captioning process. This dual attention mechanism enhances the alignment between image features and generated words, leading to more accurate and descriptive captions. The study showcases improved performance in image captioning, demonstrating the efficacy of incorporating attention mechanisms for enhancing caption generation in the context of multimedia understanding [18].

3. Methodology

3.1 Problem definition

The image (M) and its referenced statement (C) will be given, and during the model training phase, the maximum likelihood is utilised to increase the likelihood of providing an image caption as shown in Equation. 1.

$$W^* = argmax\left(\sum_{M,C} log\left(p(C|M;W)\right)\right) \quad (1)$$

Where W^* is the image caption producing model-trained parameter.

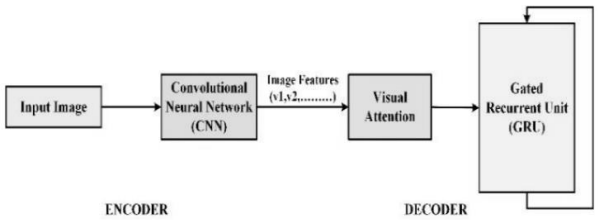

Figure. 1: Double Awareness Based Model for Image Caption Generation

Training will learn weight parameters to construct the caption C for image M. Figure 1 illustrates the model.

The attention-based technique and Gated Recurrent Units (GRU) used in image caption creation revolutionise how computers comprehend and describe visuals. The attention mechanism and GRU-based recurrent neural network are key to this process. As the caption is generated, the attention mechanism dynamically focuses on different image regions to highlight the most important sections. Dynamic attention simulates human behaviour, shifting emphasis according on caption context. For caption word prediction, the GRU-based recurrent neural network processes attended visual information and previously generated words. Captions are accurate,

coherent, and contextually relevant thanks to the GRU's sequential data modelling. These algorithms connect image details with words to provide complete and relevant captions that accurately reflect the image's visual content. It can improve content recommendation system user experiences, image indexing, and accessibility for visually impaired people. The ResNet training model is utilised in this paper as a Convolutional Neural Network. The suggested model uses visual attention techniques to produce image captions. Additionally, GRU has proven to be an effective decoder. For the given image M and descriptive sentence $S = \{S_0, S_1, S_2, \ldots, S_N\}$, the Gat ed Recurrent Model is trained, and the training process takes place as follows:

$$f_g = CNN(M) \quad (2)$$

$$x_t = Lookup(w_e, s_t), t\epsilon(0,1,2, \ldots, N) \quad (3)$$

$$h_t = GRU(x_t, v_t) \quad (4)$$

$$S_r \sim r_t = softmax(h_t) \quad (5)$$

CNN extracts global features f_g first, as showed in Equation 2. These global features activate the GRU model initially. We is the word embedding matrix for all words, and Eq. 3 shows how to find x_t, the word vector, by searching up s_t. GRU receives the current word vector x_t and visual attention vector v_t per iteration. Eq. 4 shows how it generates the concealed state. SoftMax generates selection probability vector r_t from this hidden state. Eq. 5 selects the term with the highest probability as input to the following stage. The CNN's last convolutional layer extracts v_t, the visual attention vector. At each time step, GRU receives updated vectors x_t and v_t to generate the caption word.

3.2 Long Short-Term Memory (LSTM) Vs Gated Recurrent Units (GRU)

Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRU) are two fundamental architectures in the domain of recurrent neural networks, each with its unique features and advantages. LSTMs are equipped with a more intricate structure, incorporating three distinct gates—input, output, and forget gates—that regulate the flow of information through the cell. This complexity allows LSTMs to store and manage long-term dependencies efficiently, making them well-suited for tasks where capturing extensive context is crucial. On the other hand, GRUs possess a simpler design, consolidating the reset and update functions into a single gate. This simplification results in a more streamlined architecture with fewer parameters, making GRUs computationally more efficient and faster to train. However, this simplicity also limits their ability to capture long-term dependencies effectively. As a trade-off, GRUs per-form exceptionally well in tasks where short-term dependencies are prevalent, and computational resources or training time need to be optimized. The choice between LSTM and GRU ultimately depends on the specific requirements of the task at hand, whether it necessitates a balance between memory retention and computational efficiency or places a premium on capturing extensive context for optimal performance. Gated Recurrent Units (GRU) is a type of recurrent neural network (RNN) architecture that was introduced to address the limitations of traditional RNNs, particularly the vanishing gradient problem. GRU is designed to allow for capturing long-term dependencies in data while being computationally more efficient than its predecessor, the Long Short-Term Memory (LSTM) network. GRU consists of two main gates, the reset gate and the update gate, which regulate the information flow within the net-work. The reset gate determines what information to discard from the previous state, enabling the model to focus on relevant information for the current context. The update gate decides how much of the previous state should be retained and how much of the new state should be considered.

One of the advantages of GRU is its simplified architecture, leading to fewer parameters compared to LSTM. This results in faster training times and makes GRU more computationally efficient, making it a popular choice for various applications, including natural language processing, speech recognition, and sequence-to-sequence tasks. However, GRU may face challenges in capturing very long-term dependencies compared to LSTM due to its simplified gating mechanism. Despite this, GRU has proven to be effective in a wide range of tasks and is a valuable tool in the toolkit of deep learning practitioners. The choice between GRU and LSTM often depends on the specific requirements of the task and the balance between model complexity, training efficiency, and the need to capture long-term dependencies in the data.

GRU cells track critical network data. GRU networks use two gates to do this: To reset or update the gate. The simplest GRU cell architecture is below (See Fig. 2).

A GRU cell has two inputs, as indicated below: The previous concealment, Current timestamp input. The cell mixes these and passes them through update and reset gates. To forecast the output in the current timestep, we must pass this hidden state via a thick layer with softmax activation. Thus, a new concealed state is obtained and passed to the next time step.

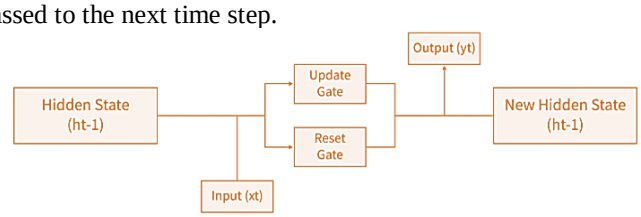

Figure. 2: The Architecture of GRU

In the context of image caption decoding using a Gated Recurrent Unit (GRU), we typically adapt the GRU equations to the specific task of generating captions for images. Let's outline the modified equations for image caption decoding with GRU:

Assuming we have an embedding of the previously generated word $y_t - 1$ and the context vector from the image c as inputs:

Reset Gate (r_t) equation—

$$r_t = \sigma(W_{ir} \cdot y_{t-1} + W_{cr} \cdot c + b_{ir} + b_{cr})$$

Update Gate (z_t) equation—

$$z_t = \sigma(W_{iz} \cdot y_{t-1} + W_{cz} \cdot c + b_{iz} + b_{cz})$$

Candidate Hidden State $(\tilde{h}_t)$ equation—

$$\tilde{h}_t = tanh(W_{in} \cdot y_{t-1} + r_t \odot (W_{hn} \cdot h_{t-1}))$$

Hidden State (h_t) Equation—

$$h_t = (1 - z_t) \odot \tilde{h}_t + z_t \odot h_{t-1}$$

Here:

- W and b are weight matrices and bias vectors associated with the GRU.
- σ is the sigmoid activation function.
- $\odot$ denotes element-wise multiplication.
- y_{t-1} is the embedding of the previously generated word at time step $t - 1$−1.
- c is the context vector from the image.
- r_t is the reset gate, controlling what information to discard.
- z_t is the update gate, determining how much of the previous state to keep.
- $\tilde{h}_t$ is the candidate hidden state, proposing an update to the hidden state.
- h_t is the hidden state at time step *t*, incorporating both word embedding's and the context vector.

In this adaptation, the GRU operates in the context of generating captions for images, where the generated word embedding's and the image context are crucial inputs influencing the captioning process.

4. Datasets

Several widely-used datasets are frequently employed in the training and assessment of image captioning algorithms. The following are a few popular image captioning datasets: The breadth, diversity, and sorts of annotations that these datasets offer allow them to be customised to meet a range of image captioning research and application goals.

5. Evaluation Metrics

By contrasting the generated captions of image captioning models with human-annotated reference captions, these measures are frequently employed to evaluate the performance of the models quantitatively. Every metric has advantages and disadvantages, therefore it's usually best to combine several for a thorough analysis.

6. Experimental Results

This section outlines the experimental details and results. We examined ResNet [12], VGG [13], DenseNet [14], GoogLeNet [15], Inceptionv4 [16], AlexNet [17], Squeezenet [1], Shufflenet [2], Mobilenet [3], MnasNet [4], NASNetLarge [5], DPN (Dual Path Network) [6], ResNext [7], WideResNet [8], SeNet [9], PolyNet [10] and InceptionResNetv2 [11] CNN models. We evaluated five ResNet (ResNet18, ResNet34, ResNet50, ResNet101, Resnet152), four DenseNet (Densenet121, Densenet169, DenseNet201, and DenseNet161), and four VGG (VGG-11, VGG-13, VGG-16, and VGG-19) models, which differ in architecture, number of parameters, accuracy, and error rates on the ImageNet dataset for object recognition.

The most popular image captioning datasets are MSCOCO, Flickr30K, and Flickr8K. Flickr30k contains 31,783 photos. MS COCO is a popular dataset for image captioning. There are 123,287 images in this collection [19].

Over 3,000,000 photos from MSCOCO are based on objects and their attributes. The Flickr30K dataset has 30,000 photos with 5 captions each. Overall, Flickr30K has 1,50,000 captions for 30K photographs. The Flickr8K dataset has 8,000 images with 5 captions. Flickr8K dataset has 30000 captions for all photographs. Other datasets include visual genome, Instagram, Stock3M, and Flickr style 10k. These datasets are efficient for image captioning across categories. Images from datasets are utilised for training, testing, and validation. Flickr8K is the model dataset in this paper.

This paper utilises Flickr8K as the model dataset. The Flickr8k dataset overview is provided in Table 1. Out of 8,000 photos, 6,000 were used for training. 1,000 photos were utilised for testing and 1,000 for validation. Captions have been pre-processed to remove words that no longer contribute to improved phrases for query images.

Various evaluation metrics are employed for image captioning. The model was tested using BLEU, METEOR, ROUGE, and CIDEr scores in this paper. The BLEU score measures the match between the generated and referenced captions. The BLEU score ranges from 0 to 1. Score 0 indicates no match between created and referred caption, while score 1 indicates complete match. Four BLEU metrics scores exist: BLEU-1, BLEU-2, BLEU-3, and BLEU-4. BLEU-1 score detects one word or gramme match between generated and referenced captions. In BLEU-2, the score indicates the match between two words or grammes in the created and referenced caption. A BLEU-3 score indicates a three-word or grammar match between the generated and referenced caption. BLEU-4 score involves matching four words or grammars between created and referred captions. We examined all ratings for the suggested model. The METEOR assessment metric focuses on synonyms of words. The BLEU score alone cannot determine the quality of created captions. In addition to these metrics, ROUGE and CIDEr were assessed and displayed in a matrix.

Table 3 demonstrates that GRU outperforms LSTM in image captioning. The suggested model was evaluated with ResNet and Inception V3 encoders, LSTM, and GRU decoders. The suggested model addresses attention mechanisms, yielding improved outcomes when ResNet and GRU are combined. Attention mechanism utilises local visual features throughout the language model process. Although global image features are conveyed once, language models incorporate local features at every word formation step.

Different ResNet, VGG, and DenseNet designs exhibit significant differences in Top-5 error on Object Detection tasks using Imagenet dataset [12-14]. This difference does not mean a corresponding difference in image captioning performance.

Table 1. Several widely-used datasets for image caption

Dataset Name	Description	Number of Images	Number of Captions per Image	Annotations
MS COCO [19]	Diverse images covering a wide range of scenes and objects, with multiple captions per image.	200,000	5	Object segmentation, keypoints, and captions.
Flickr30k [20]	31,000 images from Flickr with five descriptive captions for each image.	31,000	5	Captions
Flickr8k [21]	Smaller version of Flickr30k, containing 8,000 images with corresponding captions.	8,000	5	Captions
Visual Genome [22]	Images with scene graph annotations, providing detailed relationships between objects in the images.	108,000	Varies	Object annotations, attributes, and relationships.
ImageNet Captions [23]	Derived from the ImageNet Large Scale Visual Recognition Challenge, with captions for ImageNet images.	120,000	1	Captions
COCO-Stuff [24]	An extension of MS COCO with additional annotations for a broader rang e of object categories.	164,000	1	Object segmentation and captions.
Conceptual Captions [25]	Large-scale dataset with over 3 million images from the web, each with at least one caption.	3,300,000	1	Captions

Table 2. Evaluation Metrics used in image captioning

Metric Name	**Description**	**Range**	**Notes**
BLEU (Bilingual Evaluation Understudy) [26]	Measures n-gram overlap between generated and reference captions.	0 to 1	Higher scores indicate better agreement with reference captions. Multiple variants (BLEU-1, BLEU-2, BLEU-3, BLEU-4) are commonly used.
METEOR [27]	Combines precision, recall, and synonymy-based matching using stemming and synonym dictionaries.	0 to 1	Designed to be more language-aware and reflective of human judgment.
CIDEr [28]	Computes consensus-based similarity using cosine similarity of TF-IDF weighted n-grams.	0 to ∞	Emphasizes capturing diverse and descriptive captions.
ROUGE-L [29]	Measures overlap of longest common subsequences (LCS) between generated and reference captions.	0 to 1	Originally designed for text summarization but used in captioning.
SPICE [30]	Evaluates precision based on semantic content overlap between generated and reference captions.	0 to 1	Focuses on capturing the meaning and semantic content of captions.
WER (Word Error Rate) [31]	Measures the minimum number of word edits required to change the generated caption into the reference caption.	0 to 1	Lower scores indicate better accuracy. Commonly used in speech recognition but adapted for captioning.
METEOR-hm [27]	METEOR with stemming and synonymy-based matching but without stemming for higher-level semantics.	0 to 1	A variant of METEOR with a modified scoring approach.
TER (Translation Edit Rate) [32]	Measures the minimum number of edits required to change the generated caption into the reference caption.	0 to 1	Similar to WER but considers multiple potential reference captions.

Table 3. Performance Of CNN+LSTM+ATTENTION Method Using Different CNN Architectures (B-1 To B-4 As BLEU Score, M As METEOR, C As CIDER, R As ROUGE, S As SPICE)

Reference + CNN Model	**B--1**	**B-2**	**B-3**	**B-4**	**M**	**C**	**R**	**S**
Iandola et al. (2016) [1] + Squeezenet	60.79	42.29	28.78	19.41	18.80	46.54	44.48	12.85
Ma et al. (2018) [2] + Shufflenet	62.36	43.87	30.42	21.00	19.18	49.01	45.00	13.50
Sandler et al. (2018) [3] + Mobilenet	63.69	45.33	31.72	21.89	19.63	55.36	46.28	14.25
Tan et al. (2019) [4] + MnasNet	63.99	45.75	32.11	22.36	19.78	54.84	46.17	14.02
Zoph et al. (2018) [5] + NASNetLarge	63.60	44.66	30.16	19.93	19.73	51.34	45.49	14.00
Chen et al. (2017) [6] + DPN131	62.68	44.17	30.47	20.53	19.41	49.98	45.51	13.95
Xie et al. (2017) [7] + ResNext101	64.78	46.07	32.36	24.45	20.93	57.67	40.04	15.28
Zagoruyko et al. (2016) [8] + WideResNet101	63.47	45.37	31.71	21.73	19.84	54.27	46.23	14.51
Hu et al. (2018) [9] + SeNet154	64.23	45.94	32.54	22.62	20.81	58.45	46.83	15.05
Zhang et al.(2017) [10] + PolyNet	62.56	44.78	31.16	21.48	19.75	53.38	45.96	13.81
Szegedy et al. (2016) [11] + InceptionResNetv2	61.46	42.98	29.20	19.84	19.20	49.83	44.44	13.81
He et al. (2016) [12] + ResNet18	63.26	44.87	31.07	21.24	20.08	52.44	45.84	13.75
He et al. (2016) [12] + Resnet34	63.36	45.28	31.88	22.23	19.88	55.35	46.17	14.40
He et al. (2016) [12] + Resnet50	65.32	46.92	32.81	22.58	20.87	57.12	46.95	14.90
He et al. (2016) [12] + Resnet101	64.33	45.99	32.13	22.02	20.29	56.09	46.58	14.80

He et al. (2016) [12] + ResNet152	65.26	47.55	33.72	23.67	20.94	58.33	47.54	15.18
Simonyan et al.(2014) [13] + VGG-11	63.00	44.66	31.18	21.68	19.79	52.24	46.42	14.08
Simonyan et al.(2014) [13] + VGG-13	63.64	45.09	31.26	21.41	20.25	55.17	46.35	14.64
Simonyan et al.(2014) [13] + VGG-16	63.81	45.77	32.35	22.55	20.19	55.13	46.72	14.49
Simonyan et al.(2014) [13] + VGG-19	62.57	44.63	30.97	21.44	19.76	54.10	46.23	14.44
Huang et al. (2017) [14] + Densenet121	64.11	45.67	31.76	22.07	20.43	55.85	46.74	14.91
Huang et al. (2017) [14] + Densenet161	65.00	46.99	32.83	22.56	20.44	56.74	47.57	14.93
Huang et al. (2017) [14] + Densenet169	64.48	46.17	32.28	22.30	20.81	56.25	46.82	14.93
Huang et al. (2017) [14] + DenseNet201	64.38	46.26	32.41	22.49	20.73	59.71	47.19	15.13
Szegedy et al. (2015) [15] + GoogLeNet	62.91	44.27	30.27	20.50	19.51	50.72	46.02	13.80
Ioffe et al. (2015) [16] + Inceptionv4	60.17	42.24	28.71	19.35	18.76	48.00	44.33	13.26
Krizhevsky et al. (2012) [17] + AlexNet	59.93	40.97	27.80	19.06	18.67	46.11	44.09	12.57
Liu et al. (2018) [18] + InceptionV3 using LSTM	72.5	59.6	48.8	40.2	34.6	59.7	110.7	--
Liu et al. (2018) [18] + InceptionV4 using LSTM	72.1	59.0	48.3	39.8	34.4	59.2	113.1	--
Proposed Model (ResNet & GRU along with Attention Mechanism)	74.2	61.8	50.8	41.9	37.2	60.5	110.9	--
An Attention Model (InceptionV3 & GRU)	73.1	60.1	49.6	41.5	36.1	60.1	109.8	--

While most models create reasonable captions for images, the phrases supplied by different algorithms vary greatly. Captions generated by different algorithms may describe specific parts of an image or focus on a specific

object rather than offering a general visual summary. Sometimes, models fail to recognize certain items in images. A significant number of incidents of inaccurate gender identification suggest a statistical bias towards a specific gender in a specific setting.

The choice of CNN for the encoder considerably impacts model performance. In addition to broad observations, we may draw specific conclusions about CNN selection using ResNet [12] and DenseNet [14]. CNN architectures are ideal for image caption generation due to their decreased model complexity and superior results.

In this research, experiments were conducted utilizing Flickr8K dataset. The attention-based model employs InceptionV3 as the encoder and GRU as the decoder. The BLEU scores in this model are 0.731, 0.601, 0.496, and 0.415. Another attention-based model uses ResNet as encoder and GRU as decoder, yielding BLEU scores of 0.742, 0.618, 0.508, and 0.419. Overall, the suggested model using ResNet as encoder and GRU as decoder yielded the best results.

Figure. 3: Image Caption Generation using Proposed Model

6.1 Availability of Data and Systems

In all of the experiments, the Intel(R) Core(TM) i7-8700 CPU @ 3.20GHz, 3.19 GHz, RAM 8.00 GB (7.82 GB usable) and 64-bit operating system, x64-based processor was employed. This model was created on a Windows 10 Pro. The development process was completed using PyTorch. Dataset collected from the authorized website of MS-COCO.

7. Conclusion

This paper introduces automatic image captioning using an attention mechanism and a Gated Recurrent Unit (GRU)-based deep learning model. These components improve textual descriptions by solving the problem of creating meaningful and contextually relevant image captions. The attention mechanism dynamically aligns visual attributes with words during caption production by focusing on prominent regions. This dynamic attention helps the model describe images more accurately and contextually by emphasizing relevant details. The GRU-based system also captures word relationships in captions, modelling natural language sequentially. This generates meaningful, structured sentences, improving caption language. We proved our model's efficacy through extensive trials and conventional measurements. The resulting captions outperform baseline models in

correctness, relevance, and coherence. Captions that better explain image content have been produced by combining the attention mechanism with GRU architecture. This study has promising implications for image interpretation, accessibility, content suggestion, and more. Future study may explore multi-modal techniques, semantic context, and attention mechanisms to improve automatic image captioning and caption quality.

References

[1]. Iandola, Forrest N., Song Han, Matthew W. Moskewicz, Khalid Ashraf, William J. Dally, and Kurt Keutzer. ”SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and¡ 0.5 MB model size.” arXiv preprint arXiv:1602.07360 (2016).

[2]. Ma, Ningning, Xiangyu Zhang, Hai-Tao Zheng, and Jian Sun. ”Shufflenet v2: Practical guidelines for efficient cnn architecture design.” In Proceedings of the European conference on computer vision (ECCV), pp. 116-131. 2018.

[3]. Sandler, Mark, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. ”Mobilenetv2: Inverted residuals and linear bottlenecks.” In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4510-4520. 2018.

[4]. Tan, Mingxing, Bo Chen, Ruoming Pang, Vijay Vasudevan, Mark Sandler, Andrew How-ard, and Quoc V. Le. ”Mnasnet: Platform-aware neural architecture search for mobile.” In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2820-2828. 2019.

[5]. Zoph, Barret, Vijay Vasudevan, Jonathon Shlens, and Quoc V. Le. ”Learning transferable architectures for scalable image recognition.” In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 8697-8710. 2018.

[6]. Chen, Yunpeng, Jianan Li, Huaxin Xiao, Xiaojie Jin, Shuicheng Yan, and Jiashi Feng. ”Du-al path networks.” In Advances in neural information processing systems, pp. 4467-4475. 2017.

[7]. Xie, Saining, Ross Girshick, Piotr Dollar, Zhuowen Tu, and Kaiming ´ He. ”Aggregated re-sidual transformations for deep neural networks.” In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1492-1500. 2017.

[8]. Zagoruyko, Sergey, and Nikos Komodakis. ”Wide residual networks.” arXiv preprint arXiv:1605.07146 (2016).

[9]. Hu, Jie, Li Shen, and Gang Sun. ”Squeeze-and-excitation networks.” In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 7132-7141. 2018

[10]. Zhang, Xingcheng, Zhizhong Li, Chen Change Loy, and Dahua Lin. ”Polynet: A pursuit of structural diversity in very deep networks.” In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 718-726. 2017.

[11]. Szegedy, Christian, Sergey Ioffe, Vincent Vanhoucke, and Alex Alemi. ”Inception-v4, in-ception-resnet and the impact of residual connections on learning.” arXiv preprint arXiv:1602.07261 (2016).

[12]. He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. ”Deep residual learning for im-age recognition.” In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770-778. 2016.

[13]. Simonyan, Karen and Zisserman, Andrew. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556, 2014.

[14]. Huang, Gao, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q. Weinberger. ”Densely connected convolutional networks.” In Proceedings of the IEEE conference on computer vi-sion and pattern recognition, pp. 4700-4708. 2017.

[15]. Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Angue-lov, Dumitru Erhan, Vincent Vanhoucke, Andrew Rabinovich. Going deeper with convolu-tions. arXiv preprint arXiv:1409.4842, 2014.

[16]. Sergey Ioffe, Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In arXiv:1502.03167, 2015

[17]. Krizhevsky, Alex, Sutskever, Ilya, and Hinton, Geoffrey E. Imagenet classification with deep convolutional neural networks. In NIPS, pp. 1097–1105, 2012.

[18]. Maofu Liu, Lingjun Li, Huijun Hu, Weili Guan, Jing Tian, Image Caption Generation with Dual Attention Mechanism, Inf. Process. Manag. 57(2) (2020) 102178. doi: 10.1016/j.ipm.2019.102178.

[19]. Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., ... & Zitnick, C. L. (2014). Microsoft COCO: Common Objects in Context. In Proceedings of the European Conference on Computer Vision (ECCV), (pp. 740-755).

[20]. Young, P., Lai, A., Hodosh, M., & Hockenmaier, J. (2014). From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. Transactions of the Association for Computational Linguistics, 2, 67-78.

[21]. Rashtchian, C., Young, P., Hodosh, M., & Hockenmaier, J. (2010). Collecting image anno-tations using Amazon’s Mechanical Turk. In Proceedings of the NAACL HLT 2010 Work-shop on Creating Speech and Language Data with Amazon’s Mechanical Turk, (pp. 139-147).

[22]. Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., ... & Fei-Fei, L. (2017). Visual genome: Connecting language and vision using crowdsourced dense image annota-tions. International Journal of Computer Vision, 123(1), 32-73.

[23]. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., ... & Berg, A. C. (2015). ImageNet large scale visual recognition challenge. International Journal of Comput-er Vision, 115(3), 211-252.

[24]. Caesar, H., Uijlings, J., & Ferrari, V. (2018). COCO-Stuff: Thing and stuff classes in con-text. In Proceedings of the IEEE Conference on Computer

Vision and Pattern Recognition (CVPR), (pp. 1209-1218).

[25]. Sharma, P., Ding, N., Goodman, S., Soricut, R. (2018). Conceptual Captions: A Cleaned, Hypernymed, Image Alt-text Dataset For Automatic Image Captioning. arXiv preprint arXiv:1806.06357.

[26]. Papineni, K., Roukos, S., Ward, T., & Zhu, W. (2002). BLEU: A method for automatic evaluation of machine translation. In Proceedings of the 40th Annual Meeting of the Associa-tion for Computational Linguistics (ACL), (pp. 311-318).

[27]. Banerjee, S., & Lavie, A. (2005). METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In Proceedings of the ACL Workshop on In-trinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, (pp. 65-72).

[28]. Vedantam, R., Zitnick, C. L., & Parikh, D. (2015). CIDEr: Consensus-based image descrip-tion evaluation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), (pp. 4566-4575).

[29]. Lin, C. (2004). ROUGE: A package for automatic evaluation of summaries. In Text Summa-rization Branches Out, (pp. 74-81).

[30]. Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., & Zhang, L. (2016). SPICE: Semantic propositional image caption evaluation. In Proceedings of the European Conference on Computer Vision (ECCV), (pp. 382-398).

[31]. Martin, J., & Doddington, G. (1997). The DET curve in assessment of detection task per-formance. In Proceedings of the European Conference on Speech Communication and Technology (EUROSPEECH), (Vol. 2, pp. 1895-1898).

[32]. Snover, M., Dorr, B., Schwartz, R., Micciulla, L., & Makhoul, J. (2006). A study of transla-tion edit rate with targeted human annotation. In Proceedings of Association for Machine Translation in the Americas (AMTA), (pp. 223-231).