A HYBRID DEEP LEARNING MODEL FOR IMAGE CAPTIONING

2024
MASTER THESIS
COMPUTER ENGINEERING

Zainab Khalid TAWFEEQ

Thesis Advisor
Assist. Prof. Dr. Nehad T. A. RAMAHA

A HYBRID DEEP LEARNING MODEL FOR IMAGE CAPTIONING

Zainab Khalid TAWFEEQ

Thesis Advisor
Assist. Prof. Dr. Nehad T. A. RAMAHA

T.C.
Karabuk University
Institute of Graduate Programs
Department of Computer Engineering

Prepared as
Master Thesis

KARABUK
February 2024

I certify that in my opinion the thesis submitted by Zainab Khalid TAWFEEQ titled "A HYBRID DEEP LEARNING MODEL FOR IMAGE CAPTIONING" is fully adequate in scope and quality as a thesis for the degree of Master of Science.

Assist. Prof. Dr. Nehad T. A. RAMAHA ....

Thesis Advisor, Department of Computer Engineering

This thesis is accepted by the examining committee with a unanimous vote in the Department of Computer Engineering as a Master of Science thesis. February 05, 2024

Examining Committee Members (Institutions)		Signature
Chairman	: Assist.Prof.Dr. Nehad T. A RAMAHA (KBU)	....
Member	: Assist. Prof. Dr. İsa AVCI (KBU)	....
Member	: Assist. Prof. Dr. Ali HAMİTOĞLU (IU)	....

The degree of Master of Science by the thesis submitted is approved by the Administrative Board of the Institute of Graduate Programs, Karabuk University.

Assoc. Prof. Dr. Zeynep ÖZCAN ....

Director of the Institute of Graduate Programs

"I declare that all the information within this thesis has been gathered and presented in accordance with academic regulations and ethical principles and I have according to the requirements of these regulations and principles cited all those which do not originate in this work as well."

Zainab Khalid TAWFEEQ

ABSTRACT

M. Sc. Thesis

A HYBRID DEEP LEARNING MODEL FOR IMAGE CAPTIONING

Zainab Khalid TAWFEEQ

Karabuk University
Institute of Graduate Programs
Department of Computer Engineering

Thesis Advisor:
Assist. Prof. Dr. Nehad T. A. RAMAHA
February 2024, 76 pages

Image captioning is considered one of the most challenging tasks in computer vision. The ability of deep learning to process large amounts of visual data has played a crucial role in effectively tackling the problem of image captioning. Many studies have been introduced in this field and still need more investigation and improvements. This thesis presents a comprehensive and detailed study of the image captioning models. The study suggests utilizing various lightweight image and language models to achieve high performance in a low computational time since the image captioning process requires more time than other computer vision tasks. In this study, the Flickr30K dataset, which comprises both images and five descriptive sentences per image, is utilized. The images and the description sentences were preliminarily preprocessed to fit the next steps. Specifically, the images were resized to fit the specific dimensional requirements of the utilized models. The pre-trained models proposed in the current study include VGG-16, MobileNet, InceptionV3, XceptionNet, and ResNet50. The last classification layers were removed from all these models to get only the final feature vectors. Various lightweight models were also proposed for the language part, including LSTM, BiLSTM, GRU, and GRU with attention layers. The captions (description sentences) were preprocessed, involving cleaning, splitting, padding, and filtering, and were then provided along with the image features to the decoder part. In some training scenarios, the image and caption features are concatenated without fusion, while feature fusion was employed for others to improve the performance. Attention layers were added to focus more specifically on certain parts of the images and captions. In the experimental part, 13 training scenarios were performed. The experiments revealed that the best models with the highest performance were achieved by VGG+GRU, VGG+GRU with Attention, VGG+GRU with Feature Fusion, and MobileNet+GRU. In some experiments, the vocabulary is filtered. The algorithm selected the 15000 most frequently used phrases from the entire vocabulary to prevent it from overfitting, and this method was compared with the use of the full vocabulary. The models were evaluated using BLEU-1, BLEU-2, ROUGE, METEOR, and CIDEr metrics. The experiments conducted on the Flickr30k dataset, employing our proposed methodologies, resulted in a high BLEU-1 score of 0.674. The study was also compared with related state-of-the-art research in the same field, and the comparison proved the efficiency and high performance of the current study. The main contribution of the current study is that it introduces a comprehensive study of various image captioning models with a specific concentration on lightweight-efficient models that reduces computational time while maintaining robust performance. The study also introduces 13 various scenarios with different feature fusions and attention mechanisms to define the optimal image-textual combination for efficient, lightweight models. The findings demonstrate high performance compared to other state-of-the-art research in the same field, especially in terms of computational efficiency.

Key Words **:** Image Captioning, Image Description, Deep Learning, Image Models, Language Models, Flickr30K.

Science Code : 92432

ÖZET

Yüksek Lisans Tezi

GÖRÜNTÜ ALTYAZILAMA İÇİN HİBRİT DERİN ÖĞRENME MODELİ

Zainab Khalid TAWFEEQ

Karabük Üniversitesi
Lisansüstü Eğitim Enstitüsü
Bilgisayar Mühendisliği Anabilim Dalı

Tez Danışmanı:
Dr. Öğr. Üyesi Nehad T.A. RAMAHA
February 2024, 76 sayfa

Görüntü alt yazılanma, bilgisayarlı görü alanındaki en zahmetli görevlerden biri olarak kabul edilmektedir. Derin öğrenmenin büyük miktarda görsel veriyi işleyebilme yeteneği, görüntü alt yazılanma problemine etkin bir şekilde yaklaşmada önemli bir rol oynamaktadır. Bu alanda birçok çalışma yapılmış olup daha fazla araştırma ve iyileştirme ihtiyacı bulunmaktadır. Bu tez, görüntü alt yazılanma modelleri üzerine kapsamlı ve detaylı bir çalışma sunmaktadır. Çalışma, görüntü alt yazılanma sürecinin diğer bilgisayarlı görü görevlerine kıyasla daha fazla zaman gerektirmesi nedeniyle, düşük hesaplama süresinde yüksek performans sağlamak için çeşitli hafif görüntü ve dil modellerinin kullanılmasını önermektedir. Bu çalışmada, her bir görüntü için beş tanımlayıcı cümle içeren Flickr30K veri seti kullanılmıştır. Görüntüler ve açıklama cümleleri, sonraki adımlara uygun hale getirilmek üzere ön işlemden geçirilmiştir. Özellikle görüntüler, kullanılan modellerin belirli boyut gereksinimlerine uyacak şekilde yeniden boyutlandırılmıştır. Bu çalışmada önerilen önceden eğitilmiş modeller arasında VGG-16, MobileNet, InceptionV3, XceptionNet ve ResNet50 bulunmaktadır.

Bu modellerin son sınıflandırma katmanları kaldırılarak sadece nihai özellik vektörleri elde edilmiştir. Dil bölümü için LSTM, BiLSTM, GRU ve dikkat katmanlarına sahip GRU gibi çeşitli hafif modeller de önerilmiştir. Altyazılar (açıklama cümleleri) temizleme, bölme, doldurma ve filtreleme işlemlerinden geçirilerek ön işlemden sonra, görüntü özellikleriyle birlikte kod çözücü (Decoder) kısma sunulmuştur. Bazı eğitim senaryolarında, görüntü ve altyazı özellikleri füzyonsuz birleştirilirken, diğerlerinde performansı artırmak için özellik füzyonu kullanılmıştır. Görüntü ve altyazıların belirli kısımlarına daha özel olarak odaklanmak için dikkat katmanları (Attention layers) eklenmiştir. Deneysel bölümde, 13 eğitim senaryosu gerçekleştirilmiştir. Deneyler, en yüksek performansa sahip en iyi modellerin VGG+GRU, dikkat katmanlı VGG+GRU, özellik füzyonlu VGG+GRU ve MobileNet+GRU tarafından elde edildiğini ortaya koymuştur. Bazı deneylerde kelime hazinesi filtrelenmiştir. Algoritma, aşırı öğrenmeyi önlemek için tüm kelime dağarcığından en sık kullanılan 15.000 ifadeyi seçmiş ve bu yöntem, tam kelime haznesinin kullanımı ile karşılaştırılmıştır. Modeller, BLEU-1, BLEU-2, ROUGE, METEOR ve CIDEr metrikleri kullanılarak değerlendirilmiştir. Flickr30k veri seti üzerinde gerçekleştirilen deneyler, önerilen metodolojilerimiz kullanılarak 0.674 yüksek BLEU-1 puanı elde edilmiştir. Çalışma ayrıca, aynı alandaki ilgili güncel araştırmalarla karşılaştırılmıştır ve bu karşılaştırma, mevcut çalışmanın verimliliğini ve yüksek performansını kanıtlamıştır. Bu çalışmanın temel katkısı, hesaplama süresini azaltırken güçlü performansı koruyan hafif-etkin modellere özel bir odaklanmayla çeşitli görüntü etiketleme modellerinin kapsamlı bir çalışmasını sunmasıdır. Çalışma ayrıca etkin, hafif modeller için optimal görsel-metinsel kombinasyonu tanımlamak amacıyla farklı özellik füzyonları ve dikkat mekanizmaları içeren 13 çeşitli senaryoyu tanıtmaktadır. Bulgular, özellikle hesaplama verimliliği açısından, aynı alandaki diğer güncel araştırmalara kıyasla yüksek performans göstermektedir

Anahtar Kelimeler : Görüntü Altyazılanma, Görüntü Tanımı, Derin Öğrenme, Görüntü Modelleri, Dil Modelleri, Flickr30K.

Bilim Kodu **:** 92432

ACKNOWLEDGMENT

First, I would like to thank Allah the Almighty for His divine guidance throughout this academic journey. I am also grateful to my advisor, Asst. Prof. Dr. Nehad T.A. RAMAHA, for his support and guidance in successful completion of this thesis. My sincere thanks expand to the Karabuk University members, who played an essential role in my academic life. I am also thankful to my parents for their unwavering support. I also extend my gratitude to my colleague, Mr. Abbadullah .H Saleh, for his insightful advice and the generous sharing of his expertise. Finally, this thesis is respectfully dedicated to my homeland Iraq and to Turkey for its hospitable embrace of my academic journey.

CONTENTS

Page

Page

LIST OF FIGURES

Page

LIST OF TABLES

Page

ABBREVIATIONS

BLEU	: Bilingual Evaluation Understudy
BP	: Brevity penalty
CIDEr	: Consensus-based Image Description Evaluation
CNN	: Convolutional Neural Networks
DL	: Deep Learning
GAN	: Generative adversarial networks
GPT	: Generative pretrained transformer
GRU	: Gated Recurrent Unit
LSTM	: Long-short term memory
METEOR	: Metric for Evaluation of Translation with Explicit Ordering
NLP	: Natural language processing
Relu	: Rectified linear unit
ResNets	: Residual Networks
RNN	: Recurrent Neural Networks
ROUGE	: Recall-Oriented Understudy for Gisting Evaluation
SPICE	: Semantic propositional image caption evaluation
VGG	: Visual Geometry Group

PART 1

INTRODUCTION

1.1. OVERVIEW

Natural language processing NLP and computer vision studies have recently been more interested in the problem of automatically generating descriptive words for pictures [1] [2] [3]. The crucial duty of creating captions for photos calls for a semantic understanding of the visuals as well as the ability to craft precise and accurate description sentences. Images are one of the most readily available data kinds on the Internet in the big-data era; hence, the necessity for tagging and annotating them has grown. Because they concentrate on huge data volumes, image captioning systems are an example of a big data challenge [4]. The field of computer vision has witnessed significant interest from researchers in the past decade, particularly in the challenging domain of image captioning [2]. The primary goal of image captioning is to provide a textual description of the content depicted in an image using natural language. This task necessitates the collaborative utilization of computer vision and NLP, wherein image components are analyzed and subsequently described in a manner that resembles human language [2, 5]. Many applications involve image captioning, such as context indexing, social media content creation, education interactive learning, autonomous driving, and impaired people software (scene description with audible voice) [6, 7].

1.2. RESEARCH MOTIVATION

The previous state of the art in the field of image captioning introduced different captioning models. However, our study is the first one that experiments different combinations of visual and language models. This study considers lightweight models with different levels of features (low-level and high-level features) to see the effect of these different feature extraction models on the performance of the image captioning process and select the best one. The study also utilized different language models, starting from traditional ones with high computational time (like long-short term memory LSTM and BiLSTM) to those with low computational time and better dealing with image features (Gated Recurrent Unit GRU and feature fusion GRU).

1.3. PROBLEM STATEMENT

Previous works have made significant progress with the advent of deep learning. However, there is still a chance for improvement in generating accurate and semantically meaningful captions that align well with the image content.

The previous literature in image captioning has performed various combinations of visual and language models. However, a more comprehensive analysis and comparison of lightweight visual models with low computational language models is needed [33]. Additionally, the fusion of visual and language features and its impact on the overall performance of image captioning systems have not been well studied. To address these gaps, this study proposes a solution that concentrates on finding the best combination of visual and language models for image captioning. Specifically, the study experiments with lightweight visual models, including MobileNet, VGG16, InceptionNet, EfficientNet, and XceptionNet, and pairs them with low computational language models such as GRU and stacked GRU models. The proposed solution incorporates feature fusion techniques to leverage the joint information from visual and language models. Two fusion mechanisms are investigated: concatenating visual and language features and fusion within a single architecture. By combining the strengths of visual and language models, the proposed solution aims to achieve an enhanced feature representation and improve the overall capacity of image captioning models.

Furthermore, this study suggested various training scenarios using batch normalization layers and dropout layers and experimenting with different training parameters. The objective is to identify the best configuration that results in improved performance and robustness of the image captioning system. Besides that, different evaluation metrics, including BLEU, METEOR, CIDEr, and ROUGE, will be calculated in all training and evaluation scenarios in order to provide a comprehensive assessment of the captioning models' performance while most of the previous state-of-the-art focused on one or two metrics.

1.4. AIM AND OBJECTIVES

The goal of this thesis is to use recent deep learning models (visual and language models) for the aim of image captioning.

In order to meet this goal, the following objectives will be covered in this thesis:

- To improve the performance of the image captioning process by using the best combination of the best visual model with the best language model.
- To analyze different combinations of lightweight visual deep learning models with low computational language models to define the best combination achieving the best performance.
- To utilize the feature fusion of the visual and language model information to improve image captioning performance by achieving a better feature representation and increasing the image captioning model's capacity.
- To evaluate and compare performance using different image captioning performance metrics to define the best visual-language model.
- To try different enhancements in the proposed models (adding batch normalization layers, dropout layers, different training parameters, etc.) to define the best case.
- To compare the current proposed methods with the current and previous studies in image captioning..

1.5. SCOPE OF THE STUDY

The study utilizes different visual and language models to build the best image captioning system. The study will focus on the lightweight visual models (MobileNet, Visual Geometry Group (VGG16) model, InceptionNet, EfficientNet, and

XceptionNet) and fuse them with the lightweight language models (GRU, LSTM, GRU with Attention Layers, etc.). The study will use different combination methods focusing on the feature fusion mechanism by concatenating the output of visual and language models or fusion them in one architecture and comparing these different scenarios. The study will utilize a standard, well-known image captioning dataset (Flickr30k) dataset, which includes more than 30 thousand images with five description sentences per image.

1.6. STUDY CONTRIBUTION

Improved Performance: The study aims to enhance the performance of image captioning by defining the best combination of visual and language models. By systematically analyzing various combinations of lightweight visual deep learning models and low computational language models, the study aims to achieve superior performance in generating accurate and coherent image captions.

Innovative Combination: The study utilizes the fusion of visual and language models through feature fusion techniques. By investigating the concatenation of visual and language features, as well as alternative fusion architectures, the study aims to propose innovative approaches to get the joint information from both modalities.

Comparative Evaluation: The study evaluates and compares the performance of different image captioning models using various metrics. By conducting a thorough analysis, including popular metrics like BLEU, METEOR, CIDEr, and ROUGE, the study provides valuable insights into the strengths and weaknesses of different visual *language model combinations:* This comparative evaluation contributes to a deeper understanding of the effectiveness of different models for the image captioning task.

1.7. ORGANIZATION OF THESIS

The next chapters of the thesis will be introduced as follows:

Chapter 2 will include the literature review and related work of the image captioning field. In chapter 3, the proposed materials and methods will be introduced and well explained. The experiments, corresponding results and the discussion will be shown in chapter 4, while chapter 5 will include conclusion and future work.

PART 2

LITERATURE REVIEW & RELATED WORK

2.1. INTRODUCTION

Image captioning systems have recently evolved due to the development of deep learning models.

Many pieces of research have been introduced in the field of image captioning. However, they are all based on the same concept of any image captioning model, which requires two main parts: the image (visual) representation model and the language model.

The next paragraphs include the main concept of image captioning; then, the most recent related work will be introduced and discussed in detail.

2.2. IMAGE CAPTIONING STEPS

Any image captioning system consists of three general steps, which are the image representation (visual model), the visual encoding, and the language model [8].

Many deep learning architectures can be used in the feature representation step, including Convolutional Neural Networks (CNN), Residual Nets (ResNet) [9], VGG [10], EfficientNet [11], Generative adversarial networks (GAN) [12][13], MobileNet [14], etc.

The visual encoding part includes encoding the extracted features of the visual model in order to transfer them into an appropriate form to be fused or concatenated with the language model. It also aims to focus on the key features of the feature representations [15].

For the third part of the captioning system, which is the language model, many language architectures can be used, like Recurrent Neural Networks (RNN), Long-Short Term Memory (LSTM) [16], Transformer models [17], etc. The language model is trained using pairs of input-output sequences in order to predict the next word of a sentence in terms of previous words.

2.2.1. Feature Representation

Many deep-learning feature extraction models can be used for this step. VGG (VGG16 and VGG19), ResNet (ResNet50, ResNet101, etc.), GoogleNet, AlexNet, EfficientNet, etc. These models are developed by different researchers in order to extract the best hierarchical features effectively and perform some other tasks (like classification). For image captioning, the feature extraction part of these deep models is only used to generate image representation.

2.2.2. Visual Encoding

The global representation is the traditional method by which the activations of the last layers of the deep CNN model are used to get representations. However, some later studies used the probability distribution over common words in the description sentence [18].

Although this method is simple and extracts information about the entire input image, it leads to excessive information compression and a lack of granularity. Furthermore, it can be challenging to generate precise and detailed descriptions. On the other hand, the attention mechanism decides which part of feature representations will be introduced to the language model. This approach predicts the probability of observing a sentence using Equations (2.1) and (2.2) [8].

$$a_t = ALIGN(h_t, h_s) = \frac{\exp\left(score(h_t, h_s)\right)}{\sum_s \exp\left(score(h_t, h_s)\right)} \tag{2.1}$$

$$score(h_t, h_s) = \begin{cases} h_t^T h_s & dot \\ h_t^T W_a h_s & general \\ v_a^T \tanh(W_a h_t + U_a h_s) & Concat \end{cases} \quad (2.2)$$

Where a_t is the attention weight assigned to each source hidden state (h_s), while h_t is the current target hidden state. The content-based function is denoted as "score" and is given as Equation 2 illustrates, where W_a is the model's parameters. This function can be computed in three different ways (dot product (dot), general, or concatenation).

The last visual encoding type is the graph-based models, including semantic graphs, scene graphs, and hierarchical graph. The semantic graphs are developed with the graph convolutional neural networks [19, 20]. This type of graph combines semantic and spatial representations of the object into the LSTM model to generate a caption. On the other hand, the scene graph is more accurate and powerful since it generates structured semantic features of the image [21, 22]. It can connect objects, their relationships, and their properties in one image or sentence. The last type is the hierarchical graph or tree-based graph in which the image is divided into sub-regions, then the objects inside these sub-regions are detected, and finally, the relationships between the detected objects are defined. The tree's root represents the image, the leaves represent the segmented objects, while the hidden nodes denote the sub-regions. This method is considered the best to integrate the external semantic information and minimize the redundant interactions between representations.

2.2.3. Language Model

The language model is a regressive model that predicts the probability of showing the word z_t given previous words $\{z_1, z_2, \ldots, z_{t-1}\}$, and the image representations (features) X, which is acquired from the visual encoding model. This probability is denoted by $P(z_t \mid z_1, z_2, z_{3, \ldots,} z_{t-1,}X)$ and computed for sentences consisting of n words as follows [8].

$$\mathrm{P}(\mathrm{zt} \mid \mathrm{z1}, \mathrm{z2}, \mathrm{z3}, \ldots, \mathrm{zt}-1, \mathrm{X}) = \prod_{t=1}^{n} \mathrm{P}(\mathrm{zt} \mid \mathrm{z1}, \mathrm{z2}, \mathrm{z3}, \ldots, \mathrm{zt}-1, \mathrm{X}) \quad (2.3)$$

Recurrent Neural Networks (RNN) and LSTM are the most common choice of language models [23] [24]. Let's consider Z as the dictionary of words, z_t as the word generated at t time, h_t as the hidden state at time t, and the probability that the word z_t will be generated is p_t. z_t will also be passed to the input in the next time step. Figure 2.1. shows the architecture of the RNN model [25].

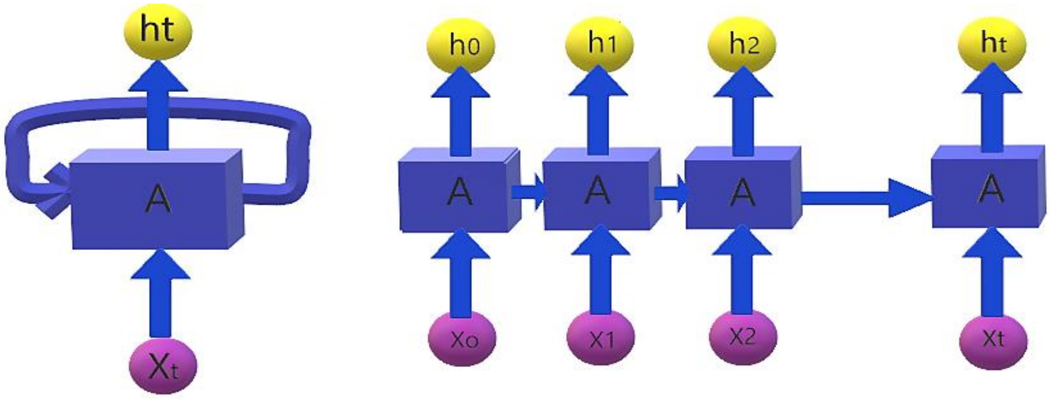


Figure 2.1. RNN architecture [23].

Where h_t is given as Equation (2.4) shows [8]:

$$h_t = RNN(h_{t-1}, X_t) \tag{2.4}$$

Where; $X_t = Ø(Z_{t-1}, \{A_i\}), \quad t>0$, $Z_t = φ(h_t, \{A_i\})$ and $X_0 = Ø_0(v) = W*v$

X_0 is the initial input of RNN and represents the result of multiplying the caption embedding by the visual features using the weight matrix W, and V. A_i represents the set of input features at time step t (visual representation of the image). The previous hidden state and the input features at time step t are combined together using the φ function to configure Z_t or the predicted word at time t. The function φ(ht, {Ai}) combines the previously hidden state h_t and other input features {Ai} to create a new representation Z_t for the current time step t. This new representation Z_t is used as input to the function Ø to compute the input for the RNN at time t (X_t).

LSTM is an advanced version of RNN that uses memory to remember and forget specific input information to predict more accurate description sentences. The main part of the LSTM model is the memory cell that is used to encode and store information about the input sequence observed up to the current time step t. LSTM cells control how much information enters the cell, stored into the cell, and outputs out of the memory cell. LSTM updates its memory cell status either by forgetting or adding information, allowing LSTM to preserve only the essential knowledge and discard the redundant data. To do this, LSTM has three gates; input, forget, and output gates. This mechanism allows LSTM to memorize contextual knowledge for the short or long term. The forget gate decides whether to memorize or discard the current value; the input gate receives the cell's input, while the output gate produces the new cell's output. The calculations of the LSTM output is given as Equations (2.5 to 2.10) illustrate [26].

$$i_i = \sigma(W_{ix}x_t + W_{ih}h_{t-1}) \tag{2.5}$$

$$f_t = \sigma(W_{fx}x_t + W_{fh}h_{t-1}) \tag{2.6}$$

$$o_t = \sigma(W_{ox}x_t + W_{oh}h_{t-1}) \tag{2.7}$$

$$c_t = f_t \odot c_{t-1} + i_t \odot h(W_{cx}x_t + W_{ch}h_{t-1}) \tag{2.8}$$

$$h_t = o_t \odot \tanh(c_t) \tag{2.9}$$

$$p_{t+1} = softmax(h_t) \tag{2.10}$$

Where; i_t, f_t, o_t, and c_t are the input, forget, outputs, and cell values at time step t, σ is the "sigmoid" activation function, x_t is the input of cell at time step t, h_t is the hidden state at time t, h_{t-1} is the previous hidden state (time t-1), W is the weight matrix (training parameters), including weights of all connections between input, forget, hidden and output cells. "Tanh" is the activation function of the output gate, while "sigmoid" is the activation function of the input and forget gates. However, the "softmax" activation function is used to compute the final probability distribution p_t over all words. Figure 2.2. illustrates the architecture of the LSTM model [26].

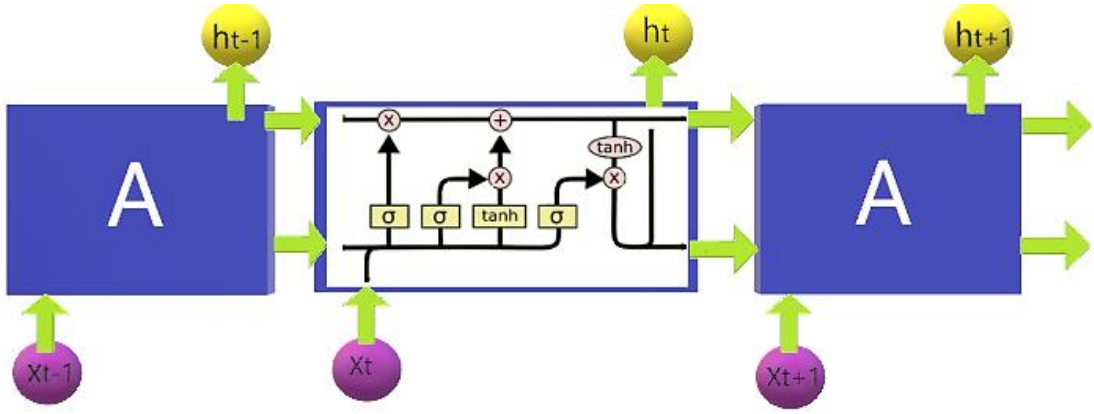


Figure 2.2. LSTM architecture.

Figure 2.3. illustrates the architecture of the CNN and LSTM-based image captioning models (in general) [27] [28].

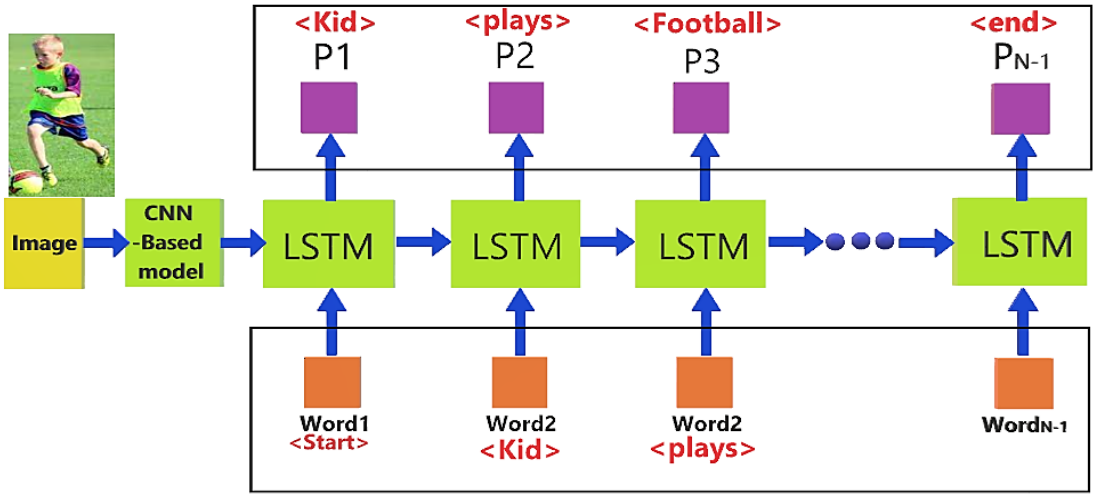


Figure 2.3. CNN and LSTM-Based image captioning models.

Transformer models [29] can process the input sequence in parallel and without need for recurrence. The transformer model consists of two main parts; an encoder and decoder.

In the encoder part. First, the input image is transformed into a sequence of representations. X= {X1, X2, …, XL}. Then, the representations are embedded using the embedding layer and passed in parallel to the encoder part of the model which consists of N identical layers, each of which contains two main parts, the self-attention layers (multi-head attention layers) and the feed-forward layer. These multi-head attention layers allow each image feature to attend to all other features of the input image, weighted by their importance for the current feature. The output of these layers are the weighted sum of the visual embedding (the importance of each visual feature). The next layer in the encoder part is the feed-forward neural network provided with a non-linear transformation function that is applied to the output of the self-attention modules. The final layer in the encoder part is the dropout layer for regularization and to avoid overfitting. The decoder is responsible for taking the input captions tokens and the output of the encoder part (importance of visual representation). The decoder also contains N identical decoding layers, each consisting of a self-attention mechanism and encoder-decoder attention mechanism. The self-attention part ensures that each token in the captioning sentence will attend to all tokens in the same sequence. The encoder-decoder part allows each word in the caption to attend to the contextual representations obtained by the encoder, weighted by their degree of importance. This step is essential to ensure that the input image and the corresponding caption are correctly aligned. Figure 2.4. shows the general architecture of the transformer-based image captioning models [30].

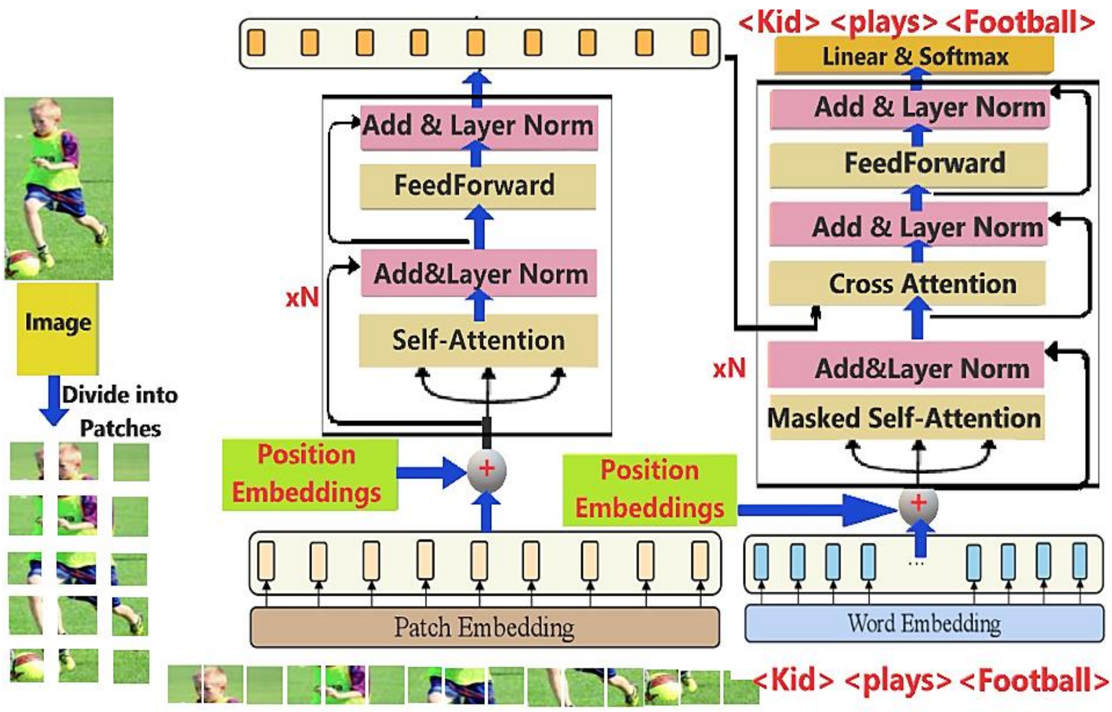


Figure 2.4. Transformer-based image captioning

2.3. RELATED WORK

In this section, many previous studies in the field of image captioning are summarized and compared. Each study will be discussed by mentioning their used methodologies, main results and their limitations.

An encoder-decoder architecture image captioning model was proposed by Sammani and Kyriazi [31]. Their system's architecture was based on an iterative refinement captioning method. This architecture consisted of two parts; the EditNet, which is a language module with a copy-LSTM model supplied with a selective copy memory attention mechanism (SCMA). DCNet was the second part of their architecture in which an LSTM architecture of de-noising auto-encoder was utilized. This main benefit of this part is to de-noise previous captions. The experiments were applied to the MSCOCO dataset with a total of 82783 training images, 40504 validation and 40775 test images. Results showed that the proposed architecture achieved a BLEU-1 of 77.9 and a BLEU-4 of 38. The CIDEr-D and SPICE values obtained by their study were 1.2 and 21.2. Their proposed methodology is time-consuming due to the refinement process.

In a research of Khan et al. [32], a multimodal architecture to perform image captioning in an end-to-end manner was introduced. Their approach involved combining a one-dimensional CNN with a pre-trained ResNet-50 model to encode sequence information. Via using this image encoder, they extracted the visual features based on regions within the images. To assess the performance of the suggested model, they employed the BanglaLekhaImageCaptions dataset, which comprised 9000 images. The assessment was conducted utilizing established metrics and a human assessment for qualitative analysis. The language model utilized in their study depends on word embedding to extract linguistic information. The experiments conducted exhibited that this approach effectively captured detailed information in the captions and generated precise and diverse captions when combined with the image features. The assessment of their approach on the chosen dataset resulted in scores of 0.651 for BLUE-1, 0.572 for CIDEr, 0.297 for METEOR, 0.434 for ROUGE, and 0.357 for SPICE. However, it is vital to note that a critical limitation of their model was that it could only recognize humans due to the constraints of the dataset utilized.

A multi-layer CNN based and LSTM image captioning approach was introduced by Poddar and Rani [33]. For the image model, they utilized the VGG16 model to extract image features, while the LSTM model was used as a language model. Many text preprocessing steps were also performed. Their experiments were applied to the Flickr8k Hindi dataset with 8000 training and 100 validation images. Their results indicated a BLEU-1 score of 0.359 and 0.55 for Unigram and Bigram, respectively. The used dataset has a moderate size. In addition, the model was evaluated using only one metric, "BLEU".

In their study, He et al. [34] introduced the image transformer architecture as a solution for image captioning tasks. They made modifications to the encoder component of the transformer model and used an implicit decoder. They utilized the R-CNN architecture to detect different parts within the image. These detected parts were then introduced to a refinement spatial visual transformer model, which consisted of three stacks. Each stack contained a multi-head dot product attention layer. The input image parts were transformed into features, namely queries, keys, and values, which were then subjected to dot product attention. The output of each multi-head attention layer was added to the input and normalized. A decoder consisting of LSTM stacks was employed to generate the descriptive sentence. The decoder took into account both the output of the encoder and the embedded features of the previously predicted word. To evaluate their model, experiments were conducted using the MSCOCO dataset, which included 113,287 images for training, 5,000 for validation, and 5,000 for testing. Words occurring less than four times were eliminated, resulting in a vocabulary of 10,369 words. Each image in the dataset was described by five sentences. The model achieved a BLEU-1 score of 81.2 and a BLEU-4 score of 39.6.

Wang et al. [35] suggested using an attention-reinforcement transformer model for image captioning. In their model, they utilized the feature attention block (FAB), which enhanced the image encoding since they detected the relationships between the image's parts. The cross-entropy and contrastive loss functions were used in the training phase. For the experimental part, they used the MSCOCO 2014 dataset (164062 images with

80% train, 10% validation, and 10% test) and the 'Karpathy' test split online server. Results showed that the proposed methodology achieved a BLEU-1 value of 81.2 and a BLEU-4 value of 39.2. The main drawback of their method is that it might add some overhead sue to the additive architectures, like the FAB block.

A spatial enhanced attention model was proposed by Hu et al. [36]. They utilized a dual spatial encoder to extract geometric correlations between image parts. The gated-normalized attention model (GNA) was also used to correct the inside attention model's distributions and reduce the redundant information and smooth gradients. All those proposed modules were applied to the original transformer model. The MSCOCO dataset was used, and the results indicated that the proposed methodology achieved a CIDEr of 134.8. Although they have good performance, their methodology requires high computational time.

The generative pre-trained transformer (GPT) was suggested by a study of Selivanov et al. [37]. In their study, they targeted image captioning in the medical domain. Two language models (GPT-3) and "Show-Attend-Tell" models were proposed in their study. The produced textual summary includes crucial details regarding the pathologies detected, their location, and 2D heatmaps that pinpoint each pathology on the scans. Three different datasets were used in the experimental part, which are the Open-I (7470 image pairs), MIMIC-CXR (377,110 images corresponding to 227,835 cases), as well as the general-purpose MSCOCO, and all images were resized into 224*224. Results showed that the proposed system achieved a BLEU-1 and BLEU-4 score of 0.725 and 0.418 on MIMIC-CXR dataset, 0.52 and 0.235 on the Open-I dataset, and 0.82 and 0.409 on the MSCOCO dataset. No state-of-art comparison between their study and others on the same medical datasets.

Fei [38] proposed an attention-aligned transformer model called "A2" for the image captioning task. His model addressed the problem of "deviated focus" in existing attention mechanisms. This model needed no annotation overhead since it was designed to guide the attention-learning process in a perturbation-based self-supervised method. His method used a mask operation on image parts in order to predict the true function of the ultimate captioning generation process. He proposed four aligned scenarios to use information (necessary image features) to refine the attention weight distribution. He applied his experiments to the MSCOCO dataset and got a BLEU-1 score of 78.6 and a BLEU-4 score of 38.2 using a Cross-entropy loss function, but by using a CIDEr score optimization, he got 81.5 and 39.8 for BLEU-1 and BLEU-4, respectively. His method's limitations included the need for manual selection of the image region features to perturb, which may not be a representative sample.

Xie et al. [39] introduced a hybrid image captioning model using Bi-LSTM and attention model. They aimed to create novel structured description sentences of the input images. Their method tried to generate sentences with a better relation to the component of the image. Besides this, they used the fast region-based CNN (Fast RCNN) architecture to detect features of image parts and objects instead of the entire image. Experiments were conducted to the Flickr30k and MSCOCO datasets. Both datasets contained five sentences describing each image. Results proved that the Bi-LS-AttM outperformed the original Bi-LSTM model in terms of BLEU score. They got 64.5 and 20.2 of BLEU-1 and BLEU-4 in the case of the Bi-LS-AttM model, while the Bi-LSTM achieved 62.1 and 19.3, respectively

2.4. IMAGE CAPTIONING DATASETS

In the first part of this section, the utilized image captioning datasets will be compared, while in the second part, the most commonly used image captioning metrics will be introduced and clarified.

Table (2-1) includes a table of the utilized dataset in the literature review studies discussed in this paper.

Table 2.1. A comparison between the used image captioning datasets.

Dataset	Number of Images	Studies	Best Result
MSCOCO	82,783 training images, 40,504 validation images, 40,775 test images 91 categories Five captions per image	Sammani & Kyriazi [31], Patwari & Naik [40], Mishra et al. [41], He et al. [34], Wang et al. [42], Castro et al. [43], Fei [36], Wang et al. [35], Parvin et al. [44], Hu et al. [36], Selivanov et al. [37], Sharma et al. [45], Yang et al. [46], Chen et al. [47], Amirian et al. [48], Deepak et al. [49], Honda et al. [50], Yan et al. [51], Chen et al. [52], Xie et al. [39]	Parvin et al. [44]: BLEU-1: 86.1
Bangla Lekha Image Captions	9000 images	Khan et al. [32]	BLUE-1: 0.651
COCO caption	330,000 images with 200,000 annotated ones 1.5 million captions Average five description sentences per image	Patwari & Naik [40]	BLEU-1: 70.6
Custom Hindi dataset based on MSCOCO	Not specified	Mishra et al. [41]	High BLEU score

Flickr8k Hindi	8,000 training images, 100 validation images	Poddar & Rani [33]	BLEU score of 0.359 (Unigram)
Flickr30k	31783 images Five description sentences per image	Padate et al. [53], Xie et al. [39]	Padate et al. [51]: BLEU-1: 65.9
MSVd	1970 video clips	Babavalian & Kiani [54]	BLEU-4: 54.82, METEOR: 35.9, Rouge: 71.6, Cider: 83.4
MSRVTT	10000 video clips	Babavalian & Kiani [54]	BLEU-4: 44.76, METEOR: 29.8, Rouge: 61.7, Cider: 52.7
MSCOCO 2014 version	328000 images	Yan et al. [51]	BLEU-1: 72.611

Table 2 shows that the most used dataset is the MSCOCO dataset, and the next most common one is the Flickr dataset. The best BLEU-1 score registered on the MSCOCO dataset is related to the study [44], with BLEU-1 equal to 86.1. The best BLEU-1 score registered on Flickr32k also corresponds to the study by Padate et al. [53] which proposed a dual attention-based model for image captioning. They started by extracting image features using the widely recognized deep learning model Inception V3. Next, they proposed a dual visual and text attention generation algorithm. This algorithm aimed to enhance the caption generation process by incorporating both visual and textual information. The final step involved generating image captions using a Bi-LSTM language model. Additionally, they employed the self-improved electric fish optimization algorithm to obtain optimal hyperparameters for the Bi-LSTM model. They conducted experiments utilizing the Flickr30k dataset. The results indicated a BLEU-1 score of 65.9 and a BLEU-4 score of 22. It is worth noting that they did not combine the visual and text attention mechanisms in the proposed model.

PART 3

MATERIALS AND METHODS

3.1. DEEP LEARNING PRINCIPLES

Deep learning is a computer-based modeling technique consisting of many processing layers used to understand the representation of data at multiple levels of abstraction. In recent years, deep learning has added image-processing opportunities to the classification process as a model for feature learning. It is an area of machine learning that accelerates approaches to reach meaningful results through detailed analysis. Image processing, video processing, and deep technology are especially popular in disciplines such as image rendering, audio analysis, biomedical signal classification, and natural language processing [55]. The most important feature of deep learning is that it works on extracting features from raw data using multiple layers to identify different relevant aspects of the input data. Deep learning techniques include convolutional, Recurrent, and deep neural networks.

3.2. THE PROPOSED DATASET

In this study, a large comprehensive and standard dataset is proposed which is Flickr30k [56]. This dataset is one of the most frequently used benchmark dataset. Each image in this dataset is descibed iusing five different sentences. It provides a large sclae of images (a total of 31783 images and almost 158915 captions).

The main reasons to choose this dataset for the aim of image captioning are:

- The dataset contains a wide range of scenes, objects, activities, and interactions helping in training a good image captioning model.

- The cpations of the dataset are written by human annotators, reflecting a realistic image understanding concept.
- The dataset includes five different captions per image allowing to explore various liguistic variations of the described image.

3.3. THE PROPOSED METHODS

The current study suggests using many lightweight DL architecutres in the visual representation part, and many low computational language models in the language model parts. The detailed methdology is shown in Figure 3.1.
In the first step of the proposed methodology, the Flickr30k dataset and its corresponding description files are acquired.

After that, both visual and language models are designed. In the visual model, the first step is the preprocessing in which the image is resized into approperiate size depending.

On the input size of model. For example, VGG16, MobileNetV2 and EfficicentNet have an input size of 224*224*3, while both Xception and InceptionV3 have an input size of 299*299. In the next step of the visual model, one of the pretrained models will be used to extract image featuures so each model will be modified by removing their classification layers and get only the final feature vector.

For the image language model, the description sentences will be first loaded and transformed into a lower-case sentence. Remove the extra spaces is the next step by which the the spaces are removed. In order to let the language model konws the start and end of each description sentence, a specific "startseq" word to start the sentence and a specific end of the sentence "endseq" to end the descrption of the image.

In the next step, all sentences are tokenized using the text tokenizer in order to transform the description sentence into a number of words (tokens). After that, the inputs and outputs of each sentence are built. Each sentence will be fed into the language model as pairs of inputs and outputs. The input starts with the "startseq" word, and the model must learn to predict the next word of the sentence.

So if the sentence is "boy plays with ball", the first start token will be "startseq" and the first output word will be "boy". In the next time step, the input word will be "boy" while the output word will be "plays" and so on until reaching the final input word which will be "ball", while the output word must be "endseq". In the next step, all sentences are padded to the specified maximum caption length. Then the one-hot encoding is performed to transform tokens into an approperiate form for the input of the language model. In the next step, the visual encoding task is performed to the feature vector of the visual model in order to transfrom the visual features into a form that can be combined with the language model.

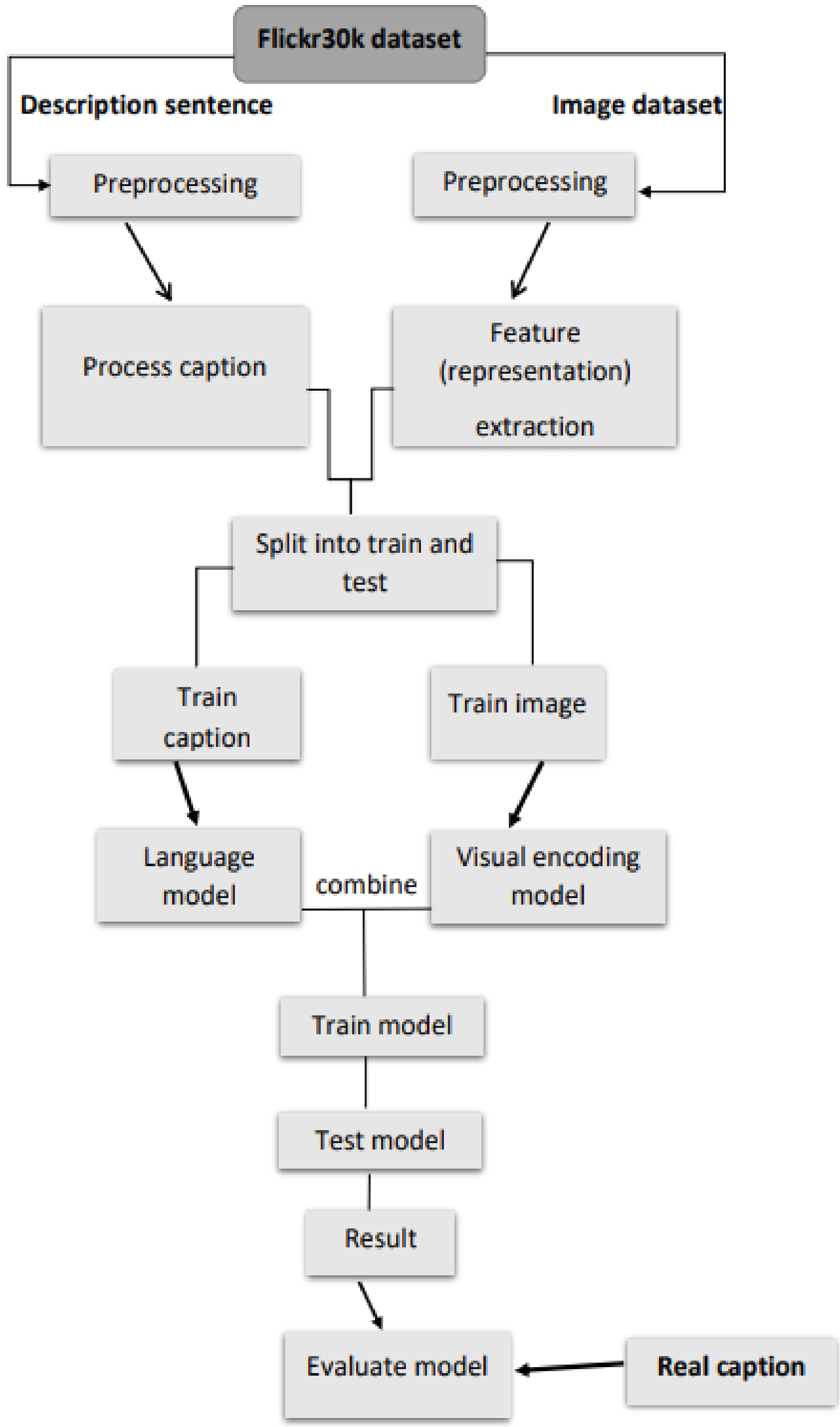


Figure 3.1. The proposed Methdologies and models

The visual encoding model consists of dropout layer of 50%, Dense layer with 512 units and ReLU activation function, which applies a linear transformation followed by a rectified linear unit (ReLU) activation function to introduce non-linearity, a batch normalization layer that normalizes the activations of the previous layer throughout batches.

The textual encoded features are introduced to the language model which consists of an embedding layer (converts the input words into dense vectors of fixed size, Dropout layer which helps in regularization by randomly setting a fraction of input units to 0 during training,), main backbone model (can be LSTM, Bi-LSTM, GRU or stacked GRU), and finally a batch normalization layer. This proposed architecture is the best one of many possible ones that are tried experimentally by us in order to define the best architecture achieving the best performance.

The visual features and the textual encoded descriptions are now ready to be introduced fusion part. Two different approaches are used in this step; in the first one the image features resulted from the visual encoder and the sequence features of the desription sentences are either added (fused in one feature vector) or concatednated (fused to constitute a bigger feature vector). In the first case, the size of both vectors must be the same, while in the second one the size of feature vectors can be different. Figure (3-2) illustrates the architecture of the visual and language fused model.

The final decoder model consists of a dense layer of 512 neurons and relu activation function, a dropout layer of percentage 50%, and a final dense layer of the size of the vocabulary. Figure (3-4) illustrates this architecture.

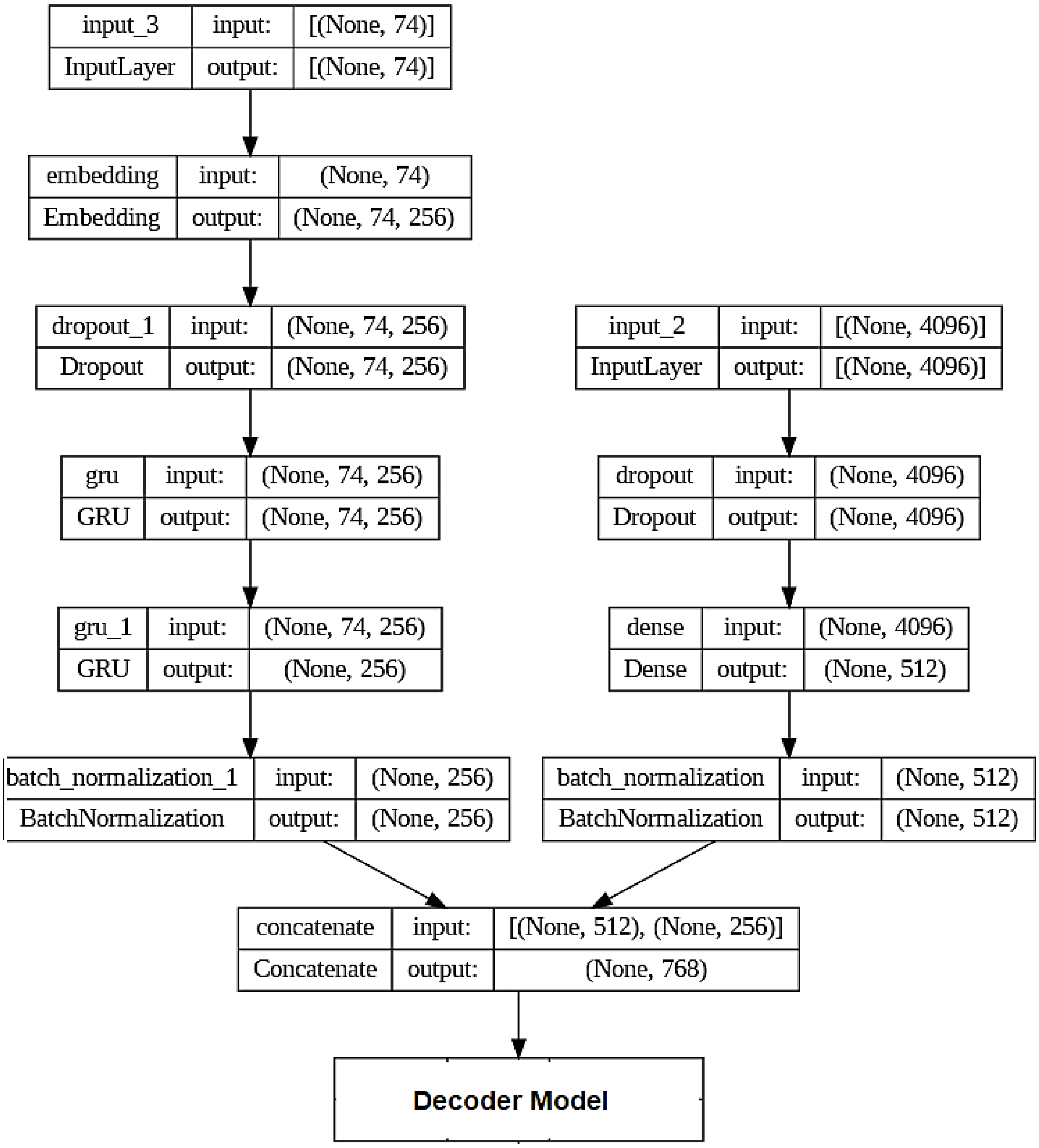


Figure 3.2. The proposed VGG-GRU-Concatenation based fusion image captioning model.

While Figure 3.3. shows the same architecture but using the "Add" fusion method in which the corresponding features of both visual and textual vectors will be fused.

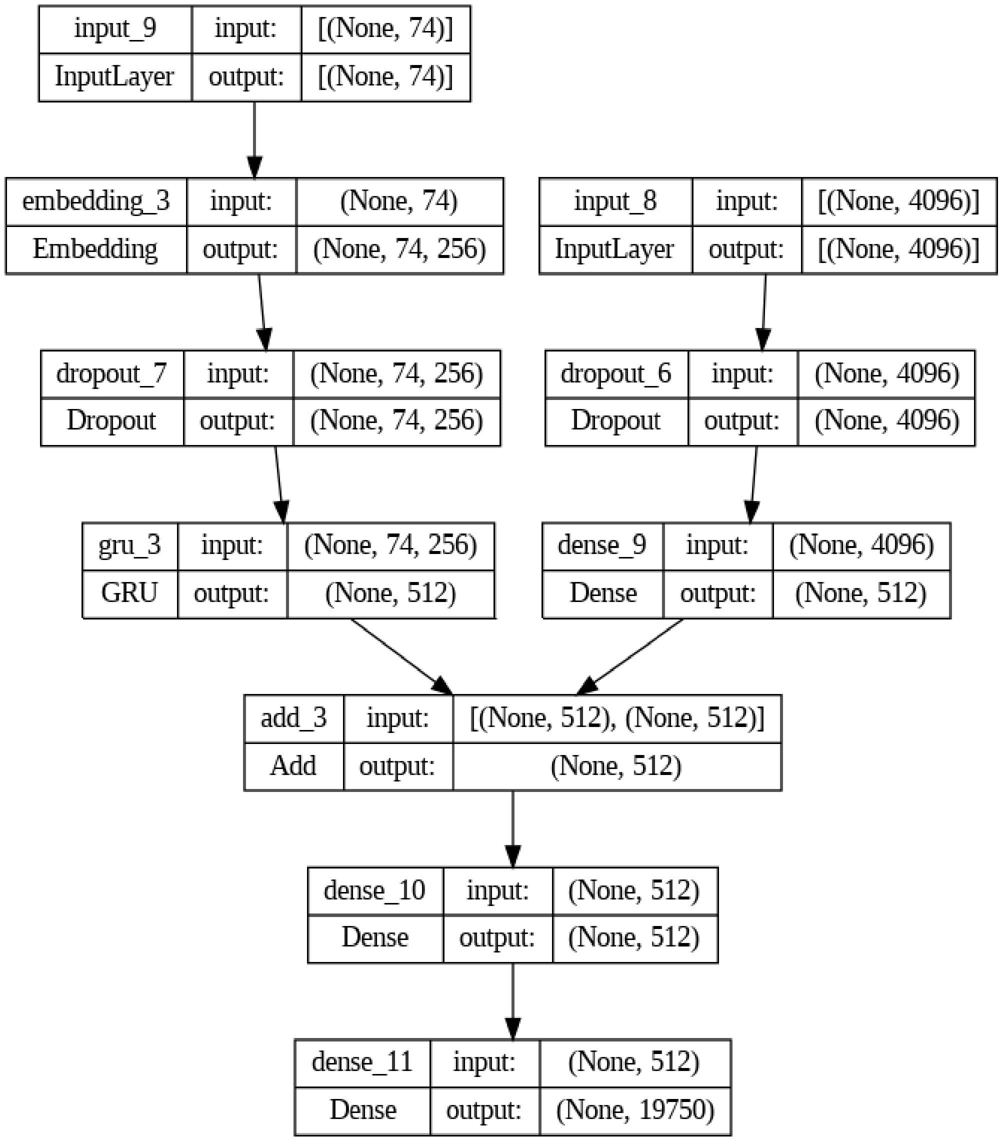


Figure 3.3. Another proposed VGG-GRU-Addition based fusion image captioning model with different architecure of model in Figure 3.2.

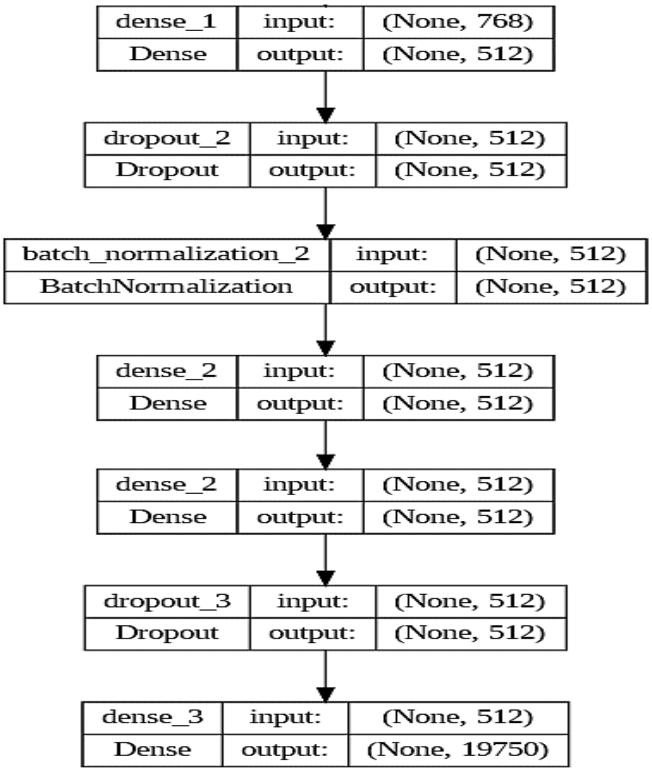


Figure 3.4. The proposed decoder model

The model is compiled using the Adam optimizer and the categorical cross entropy function since the current problem is a multi-class problem (categorical problem).

The precious architecture is the main architecture used in the experimental part. However, different architectures are experimented including various visual models and different language models.

3.4. VGG-GRU WITH ATTENTION LAYER MODEL

In this main modification of the proposed models, an attention layer is added to the language model in order to allow model focus on different part of the image when gengerating the captions will improve the perofmance of the image captioning process. Figure 3.5. shows the artictecture of the VGG-GRU attention-based model with GRU fusion in which the Attention is applied to both image and captions features.

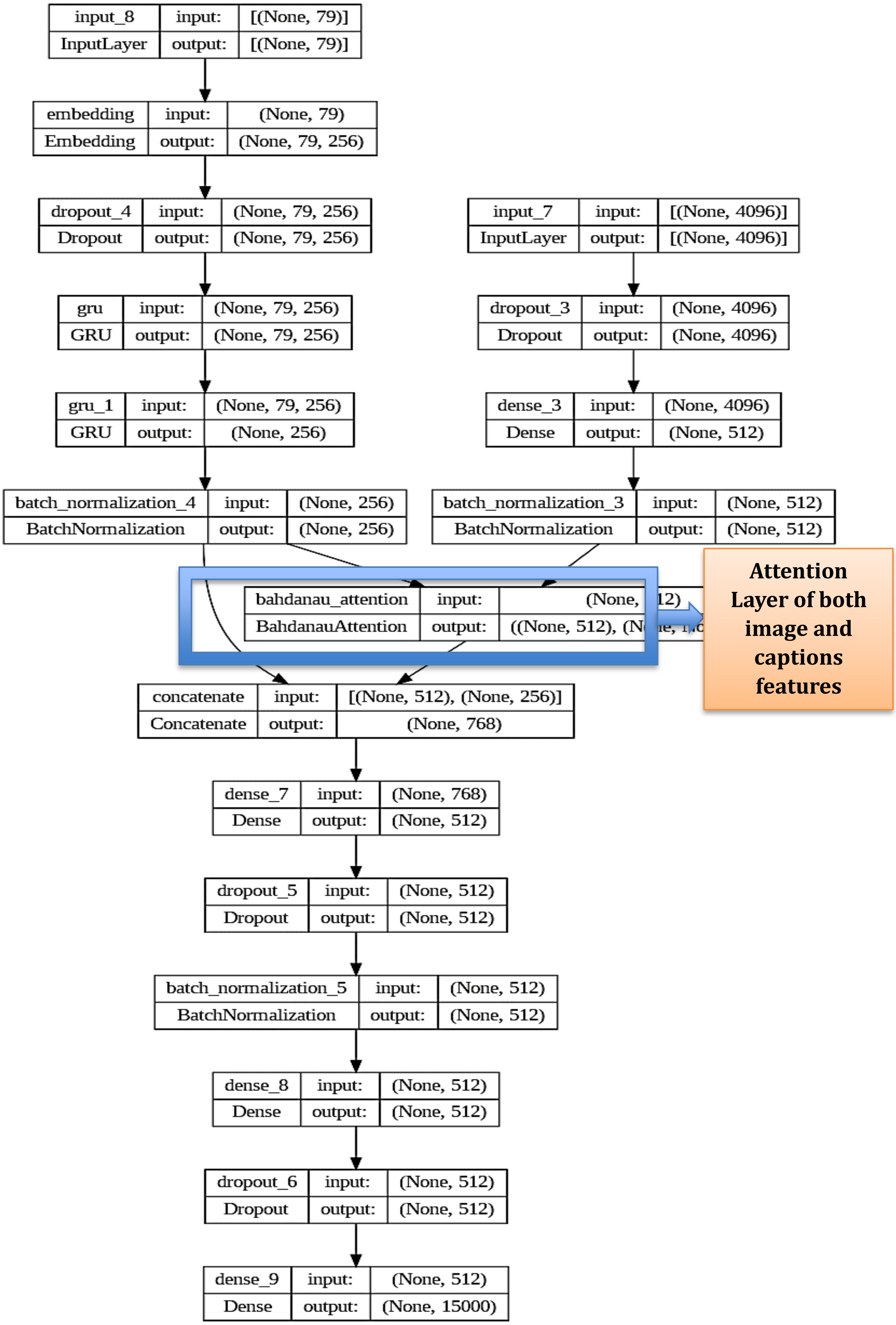


Figure 3.5. The proposed VGG-GRU attention based with Feature Fusion model

3.5. EVALUATION METRICS

For any image captioning or description system, the performance must be evaluated using specific evaluation metrics which are different from the known classification or segmentation metrics.

The image captioning evaluation metrics include the following:

- Manual evaluation: this type of image captioning evaluation refers to relevance to the source image, fluency of expression, expression variety, etc. These metrics are accurate but require too much computation [57].
- Rule-based evaluation metrics: these metrics compute the degree of correlation between the generated captions and the original description sentences and include the following metrics:

  a. Bilingual Evaluation Understudy (BLEU) [58]: this metric is commonly used in machine translation applications to compute the degree of overlap between the generated captions and the original reference description in n tuples (with n-gram where n=1,2, 3, and 4), so we can compute BLEU-1, BLEU-2, BLEU-3, and BLEU-4 metrics. The more BLEU score, the higher overlap between original and generated captions. BLEU metric problem is that it is affected by the length of the generated captions, so it will be higher if the captioning sentence is small, i.e., the higher values of BLEU do not actually mean a better description [57]. BLEU is given as Equation 3.1 shows [58].

$$p_n = \frac{\sum_{c \in candidate} \sum_{n-gram \in c} Count_{clip}(n-gram)}{\sum_{c' \in candidate} \sum_{n-gram' \in c'} Count(n-gram')} \qquad (3.1)$$

Where n-gram is the number of sequential words being n. In Equation 10, the numerator's first summation symbol Σcandidate sums all candidates, as there may be multiple sentences during calculation. The second summation Σn-gram sums all n-$gram$ in a candidate (c), where $Countclip(n\text{-}gram)$ refers to the number of occurrences of a certain n-gram in the reference caption. For the denominator, the summation symbols have the same meaning as in the numerator. $Count\ (n\text{-}gram')$ refers to the number of $n\text{-}gram'$ occurrences in the candidate. The denominator computes the number of $n\text{-}gram$ acquired from all candidates. BLEU incorporates a brevity penalty, denoted as BP, in order to prevent extremely brief translations that aim to maximize their precision scores. BLEU and BP are given as Equation 3.2 describes [58].

$$BP = \begin{Bmatrix} 1 & if\ |c| > |r| \\ e^{(1-|r|/|c|)} & if\ \ |c| \leq |r| \end{Bmatrix}$$

$$BLEU = BP\ .\exp(\textstyle\sum_{n=1}^{N} w_n log(p_n)) \tag{3.2}$$

Where |c| and |r| denote to the size of the result translation (caption) and reference translation, respectively. N is equal to 4, and w_n denotes the weighting factor, which is actually set to 1/N.

b. Consensus-based Image Description Evaluation (CIDEr) [59]: This metric was proposed based on chunks taking into account grammaticality, salience, importance, and accuracy, thereby reducing the impact of high-frequency n-grams on the results. This evaluation metric assesses the correlation between a sentence generated by an image captioning model and a set of reference sentences that are manually annotated by humans.
c. Metric for Evaluation of Translation with Explicit Ordering (METEOR) [60]: this metric is proposed to solve the problem of the effect of short sentences on the BLEU score. In this evaluation metric, the chunk is utilized as the main unit of evaluation, while the final performance evaluation is based on the F-value which is a combination of recall and accuracy scores. So, the METEOR score uses the chunking algorithm and external resources (which is different from other metrics like BLEU and CIDEr), which can cause some instability in performance. The unigram precision, unigram recall, and fragmentation measure are used in this metric to compute the final score. The purpose of this measure is to assess

the degree of coherence in the matched words of the generated description relative to the reference description. Evaluating METEOR involves examining the degree of correlation between metric scores and human judgments of the quality of the descriptions.

d. Semantic propositional image caption evaluation (SPICE) [61]: computes the correlation of the generated description with the original image-based scene graph. So, SPICE computes the ratio by which the generated captions cover the entities and inter-entity relationships in the original image. However, SPICE is similar to human judgment. However, the log-likelihood score of metrics like BLEU, METEOR, and CIDEr are less similar to human judgment. Although SPICE and CIDEr are the nearest metrics to human judgment, they are the least optimizable metrics [59].

e. Recall-Oriented Understudy for Gisting Evaluation (ROUGE) [57]: This metric is usually used to evaluate the text summarization models. ROUGE only calculates recall by considering the number of overlapping units between the predicted descriptions and the reference description tuples. Equation 3.3 illustrates the computations.

$$ROUGH = \frac{N_{overlap}}{N_{total}} \tag{3.3}$$

f. BERTScore [62]: the BERT score to resolve previous image captioning metrics. However, a different configuration of BERT scores results in different trade-offs, and they depend on the domain and used language. BERT computes the similarity between each word in the predicted sentence and the corresponding word in the reference sentence. The word similarity is computed based on contextual embedding rather than comparing words together.

PART 4

RESULTS AND DISCUSSION

4.1. INTRODUCTION

The main results obtained from the image captioning models will be introduced in this chapter. The main training scenarios and their corresponding results (with performance metrics) will also presented and concluded.

4.2. PROPOSED TRAINING SCENARIOS

In this study, 13 different training scenarios are proposed in order to define the best image captioning model among many available options. These scenarios are suggested in terms of the notes acquired from previous studies.

The proposed scenarios are based on two main changes: the image model (feature extraction model) and the language model. So, changing the image model, the language model, and their corresponding parameters or adding a modification to the language model architecture results in a new combination of the image captioning model (new scenario).

Table 4.1. includes the training parameters of all models.

Table 4.1. Training parameters of all models.

Parameter	Value
Vocabulary Size	19750
Reduced Vocabulary Size	15000
Maximum Caption Length	74
Training set percentage	80%
Test set percentage	20%
Embedding layer size	256
GRU layer size	256
Hidden layer activation function	Relu
Epochs	50
Batch Size	512
EarlyStopping condition	Number of epochs without enhancement= 25
Performance Monitor	Validation Loss
Save only best model	True
Optimizer	Adam

4.3. VGG-BASED TRAIİNING SCENARIOS

In this part, the VGG image model and many other language model options will be used. Figure 4.1. shows the architecture and number of trainable parameters of the VGG-16 model. VGG-16 model consists of 134260544 parameters, and the output feature vector is of size 4096 (after eliminating the last two layers, which are the classification layers, and we do not need them since we only need the feature vector).

The following scenarios are proposed:

- 1- VGG-16 as an image model and LSTM as a language model.
- 2- VGG-16 as an image model and Bi-LSTM as a language model.
- 3- VGG-16 as an image model and GRU as a language model.
- 4- VGG-16 as an image model and GRU Fusion as a language model.

All these scenarios are built, trained, and evaluated.

Figure 4.2 shows the results of testing the image captioning of all VGG-Based image captioning models using some of the test set samples. The predicted description of each image is illustrated side-by-side with the original five description sentences to compare both actual and predicted captioning sentences of all scenarios.

```
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 input_1 (InputLayer)        [(None, 224, 224, 3)]     0

 block1_conv1 (Conv2D)       (None, 224, 224, 64)      1792

 block1_conv2 (Conv2D)       (None, 224, 224, 64)      36928

 block1_pool (MaxPooling2D)  (None, 112, 112, 64)      0

 block2_conv1 (Conv2D)       (None, 112, 112, 128)     73856

 block2_conv2 (Conv2D)       (None, 112, 112, 128)     147584

 block2_pool (MaxPooling2D)  (None, 56, 56, 128)       0

 block3_conv1 (Conv2D)       (None, 56, 56, 256)       295168

 block3_conv2 (Conv2D)       (None, 56, 56, 256)       590080

 block3_conv3 (Conv2D)       (None, 56, 56, 256)       590080

 block3_pool (MaxPooling2D)  (None, 28, 28, 256)       0

 block4_conv1 (Conv2D)       (None, 28, 28, 512)       1180160

 block4_conv2 (Conv2D)       (None, 28, 28, 512)       2359808

 block4_conv3 (Conv2D)       (None, 28, 28, 512)       2359808

 block4_pool (MaxPooling2D)  (None, 14, 14, 512)       0

 block5_conv1 (Conv2D)       (None, 14, 14, 512)       2359808

 block5_conv2 (Conv2D)       (None, 14, 14, 512)       2359808

 block5_conv3 (Conv2D)       (None, 14, 14, 512)       2359808

 block5_pool (MaxPooling2D)  (None, 7, 7, 512)         0

 flatten (Flatten)           (None, 25088)             0

 fc1 (Dense)                 (None, 4096)              102764544

 fc2 (Dense)                 (None, 4096)              16781312

=================================================================
Total params: 134260544 (512.16 MB)
Trainable params: 134260544 (512.16 MB)
Non-trainable params: 0 (0.00 Byte)
```

Figure 4.1. VGG-16 model.

Image & Original Description	Predicted Description	
 1- child in pink dress is climbing up set of stairs in an entry way 2- little girl in pink dress going into wooden cabin 3- little girl climbing the stairs to her playhouse 4- little girl climbing into wooden playhouse 5- girl going into wooden building	**(VGG+LSTM)** girl climbing her wooden climbing set in her playhouse	**(VGG+BiLSTM)** child in pink dress is climbing up set to her playhouse
	(VGG+BiLSTM) with different training parameters man in blue shirt is sitting on the ground	**VGG+GRU** girl climbing into wooden cabin
	VGG+GRU (Feature Fusion) little girl climbing into wooden cabin	
	(VGG+LSTM) little girl is sitting in paint with pigtails painting in front of rainbow canvas with rainbow painting.	**(VGG+BiLSTM)** there is girl with pigtails sitting on the grass with her hands on it
	(VGG+BiLSTM) with different training parameters	**VGG+GRU** little girl is sitting in front of painted rainbow

<table>
<tr><th>Image & Original Description</th><th colspan="2">Predicted Description</th></tr>
<tr><td rowspan="2">
1- small girl in the grass plays with fingerpaints in front of white canvas with rainbow on it.
2- little girl covered in paint sits in front of painted rainbow with her hands in bowl
3- there is girl with pigtails sitting in front of rainbow painting
4- little girl is sitting in front of large painted rainbow
5- young girl with pigtails painting outside in the grass</td><td colspan="2">little boy in red shirt is playing on the beach</td></tr>
<tr><td colspan="2">VGG+GRU (Feature Fusion)
little girl in paint sits in painted rainbow with her hands in bowl</td></tr>
<tr><td rowspan="2"></td><td>(VGG+LSTM)
man is standing in front of skyscraper</td><td>(VGG+BiLSTM)
there is skyscraper in the distance with man walking on the distance</td></tr>
<tr><td>(VGG+BiLSTM) with different training parameters
man in red shirt and white shirt is standing on the sidewalk</td><td>(VGG+ GRU)
man stands in front of skyscraper</td></tr>
</table>

Image & Original Description	Predicted Description	
 1- there is skyscraper in the distance with man walking in front of the camera 2- behind the man in red shirt stands large skyscraper 3- man stands in front of very tall building 4- man is standing in front of skyscraper 5- man stands in front of skyscraper	**VGG+GRU (Feature Fusion)** man in blue shirt and jeans is walking through front of skyscraper	
	(VGG+LSTM) three dogs on grassy hill with three dogs in field	**(VGG+BiLSTM)** three dogs are playing on grassy hill
	(VGG+BiLSTM) with different training parameters dog is running on the beach	**(VGG+GRU)** two dogs are playing in the grass

<table>
<tr><th>Image & Original Description</th><th colspan="2">Predicted Description</th></tr>
<tr><td>
1- three dogs are standing in the grass and person is sitting next to them.
2- three dogs stand in grassy field while person kneels nearby.
3- three dogs are playing on grassy hill with blue sky
4- woman crouches near three dogs in field
5- three dogs on grassy hill</td><td colspan="2">VGG+GRU (Feature Fusion)
three dogs are playing in grassy field with cow kneels in the background</td></tr>
<tr><td rowspan="3">
1- boy in blue swimming trunks slides down yellow slide into wading pool with inflatable toys floating in the water
2- child is falling off slide onto colored balloons floating on pool of water
3- boy sliding down slide into pool with colorful tubes
4- boy in blue shorts slides down slide into pool
5- boy rides down slide into small backyard pool</td><td>(VGG+LSTM)
boy sliding down slide into pool.</td><td>(VGG+BiLSTM)
there is boy sliding down slide into pool.</td></tr>
<tr><td>(VGG+BiLSTM) with different training parameters
boy sliding into pool.</td><td>(VGG+ GRU)
two boys in swimming pool</td></tr>
<tr><td colspan="2">VGG+GRU (Feature Fusion)
boy sliding down slide into pool with colorful tubes</td></tr>
</table>

Figure 4.2. Results of testing the image captioning VGG-LSTM, VGG-BiLSTM, VGG-GRU, and VGG-GRU Feature Fusion models using some of test set samples.

All test samples are described in successful and similar descriptions compared to the original descriptions in most of the proposed models. However, the worst model is the VGG-LSTM with a BLEU-1 score of 0.45851 and bad description results, while the best model is the VGG-GRU Feature fusion model with BLEU-1 score of 0.664 and captioning results, which are very similar to the original ones, in Table 4.2. The performance evaluation results are concluded using BLEU, ROUGE, CIDEr, METEOR, and Loss.

Table 4.2 Results of training the VGG16-based image captioning models.

Model	**Loss**	**Val-Loss**	**BLEU-1**	**BLEU-2**	**ROUGE**	**CIDEr**	**METEOR**
VGG+LSTM	2.2245	6.1677	0.458519	0.203514	0.3049	0.2883	0.3162
VGG+BiLSTM	4.4897	4.8086	0.498595	0.255005	0.2995	0.1092	0.3061
VGG+BiLSTM Different training Parameter	4.4358	4.7641	0.523388	0.269940	0.3010	0.1158	0.3090
VGG+GRU	3.4589	4.19689	0.620133	0.387619	0.3278	0.2925	0.3697
VGG+GRU Fusion	4.4177	4.1833	0.663963	0.401557	0.3150	0.2383	0.3453

4.4. MOBILENET-BASED TRAINING SCENARIOS

In this part, the VGG image model and many other language model options will be used.

Figure 4.3 shows the architecture and number of trainable parameters of the MobileNet model.

The number of trainable parameters are 4231976, and the output feature vector of the models is of length 1000 (which is smaller than VGG-16).

The following MobileNet scenarios are proposed (in terms of the previous succession of the GRU language model, the GRU layers are also proposed as the main scenarios of this part).

- MobileNet as an image model and LSTM as a language model.
- MobileNet as an image model and Bi-LSTM as a language model.
- MobileNet as an image model and GRU as a language model.
- MobileNet as an image model and GRU Fusion as a language model.

All these scenarios are built, trained, and evaluated.

Figure 4.4. shows the results of testing the image captioning of all MobileNet-based image captioning models using some of the test set samples. The predicted description of each image is illustrated side-by-side with the original five description sentences to compare both actual and predicted captioning sentences of all scenarios.

```
 Layer (type)                    Output Shape              Param #
=================================================================
 input_1 (InputLayer)            [(None, 224, 224, 3)]     0

 conv1 (Conv2D)                  (None, 112, 112, 32)      864

 conv1_bn (BatchNormalizatio      (None, 112, 112, 32)     128
 n)

 conv1_relu (ReLU)               (None, 112, 112, 32)      0

 conv_dw_1 (DepthwiseConv2D)      (None, 112, 112, 32)     288

 conv_dw_1_bn (BatchNormaliz      (None, 112, 112, 32)     128
 ation)

 conv_dw_1_relu (ReLU)           (None, 112, 112, 32)      0

 conv_pw_1 (Conv2D)              (None, 112, 112, 64)      2048

 conv_pw_1_bn (BatchNormaliz      (None, 112, 112, 64)     256
 ation)

 conv_pw_1_relu (ReLU)           (None, 112, 112, 64)      0

 conv_pad_2 (ZeroPadding2D)      (None, 113, 113, 64)      0

 conv_dw_2 (DepthwiseConv2D)      (None, 56, 56, 64)       576

 conv_dw_2_bn (BatchNormaliz      (None, 56, 56, 64)       256
 ation)

 conv_dw_2_relu (ReLU)           (None, 56, 56, 64)        0

 conv_pw_2 (Conv2D)              (None, 56, 56, 128)       8192

 conv_pw_2_bn (BatchNormaliz      (None, 56, 56, 128)      512
 ation)

 conv_pw_2_relu (ReLU)           (None, 56, 56, 128)       0

 conv_dw_3 (DepthwiseConv2D)      (None, 56, 56, 128)      1152

 conv_dw_3_bn (BatchNormaliz      (None, 56, 56, 128)      512
 ation)

 conv_dw_3_relu (ReLU)           (None, 56, 56, 128)       0

 conv_pw_3 (Conv2D)              (None, 56, 56, 128)       16384

 conv_pw_3_bn (BatchNormaliz      (None, 56, 56, 128)      512
 ation)

 conv_pw_3_relu (ReLU)           (None, 56, 56, 128)       0
                                ⋮
 conv_pw_13 (Conv2D)             (None, 7, 7, 1024)        1048576

 conv_pw_13_bn (BatchNormali      (None, 7, 7, 1024)       4096
 zation)

 conv_pw_13_relu (ReLU)          (None, 7, 7, 1024)        0

 global_average_pooling2d (G      (None, 1, 1, 1024)       0
 lobalAveragePooling2D)

 dropout (Dropout)               (None, 1, 1, 1024)        0

 conv_preds (Conv2D)             (None, 1, 1, 1000)        1025000

 reshape_2 (Reshape)             (None, 1000)              0

=================================================================
Total params: 4,253,864
Trainable params: 4,231,976
Non-trainable params: 21,888
_________________________________________________________________
```

Figure 4.3. MobileNet model.

Image & Original Description	Predicted Description
 1- child in pink dress is climbing up set of stairs in an entry way 2- little girl in pink dress going into wooden cabin 3- little girl climbing the stairs to her playhouse 4- little girl climbing into wooden playhouse 5- girl going into wooden building	**(MobeileNet+BiLSTM)** little girl climbing into wooden cabin **(MobeileNet+GRU)** little girl climbing into wooden cabin **(MobeileNet+GRU Feature Fusion)** little girl climbing into wooden playhouse
 1- small girl in the grass plays with fingerpaints in front of white canvas with rainbow on it. 2- little girl covered in paint sits in front of painted rainbow with her hands in bowl 3- there is girl with pigtails sitting in front of rainbow painting 4- little girl is sitting in front of large painted rainbow 5- young girl with pigtails painting outside in the grass	**(MobeileNet+BiLSTM)** little girl is sitting in front of painted rainbow **(MobeileNet+GRU)** little girl is sitting in front of painted rainbow **(MobeileNet+GRU Feature Fusion)** little girl in pigtails is running outside of the rainbow

Image & Original Description	Predicted Description
 1- there is skyscraper in the distance with man walking in front of the camera 2- behind the man in red shirt stands large skyscraper 3- man stands in front of very tall building 4- man is standing in front of skyscraper 5- man stands in front of skyscraper	**(MobeileNet+BiLSTM)** man stands in front of skyscraper **(MobeileNet+GRU)** man stands in front of skyscraper **(MobeileNet+GRU Feature Fusion)** man stands in front of skyscraper
	(MobeileNet+BiLSTM) three dogs on grassy hill **(MobeileNet+GRU)** three dogs are walking on the grass

Image & Original Description	**Predicted Description**
 1- three dogs are standing in the grass and person is sitting next to them. 2- three dogs stand in grassy field while person kneels nearby. 3- three dogs are playing on grassy hill with blue sky 4- woman crouches near three dogs in field 5- three dogs on grassy hill	**(MobeileNet+GRU Feature Fusion)** three dogs are running through grassy field
 1- boy in blue swimming trunks slides down yellow slide into wading pool with inflatable toys floating in the water 2- child is falling off slide onto colored balloons floating on pool of water 3- boy sliding down slide into pool with colorful tubes 4- boy in blue shorts slides down slide into pool 5- boy rides down slide into small backyard pool	**(MobeileNet+BiLSTM)** boys in swimming pool **(MobeileNet+GRU)** boy sliding down slide into wading pool with inflatable toys **(MobeileNet+GRU Feature Fusion)** boy sliding down slide into wading pool with inflatable toys floating in the water

Figure 4.4. Results of testing the image captioning MobileNet-LSTM, MobileNet-GRU, and MobileNet-GRU Feature Fusion models using some of test set samples.

MobileNet-based image captioning models achieve a good performance. However, the worst model is the MobileNet-BiLSTM with a BLEU-1 score of 0.5916, but it is better than the VGG-BiLSTM model, while the best model is the MobileNet-GRU with a BLEU-1 score of 0.654 and captioning results which are very similar to the original ones. However, the VGG-GRU model achieves a better BLEU-1 score than the corresponding MobileNet-GRU model. Table 4.3. The performance evaluation results are concluded using BLEU, ROUGE, CIDEr, METEOR, and Loss of all MobileNet-based models.

Table 4.3. Results of training the MobileNet-based image captioning models.

Model	**Loss**	**Val-Loss**	**BLEU-1**	**BLEU-2**	**ROUGE**	**CIDEr**	**METEOR**
MobileNet + BiLSTM	4.6787	6.1725	0.516679	0.244216	0.2987	0.1096	0.3053
MobileNet + GRU	4.1251	4.1251	0.654240	0.352221	0.3283	0.3345	0.3593
MobileNet + GRU Feature Fusion	4.2774	4.1764	0.649498	0.380580	0.3185	0.2044	0.3313

4.5. OTHER IMAGE CAPTIONING TRAINING MODELS SCENARIOS

In this part, many other image models are utilized to define the best image model. The models utilized are XceptionNet, ResNet50, EfficientNetB0, and InceptionV3. Figure 4.5. shows the architecture of many proposed image captioning models and the number of trainable parameters.

Image & Original Description	Predicted Description
 1- child in pink dress is climbing up set of stairs in an entry way 2- little girl in pink dress going into wooden cabin 3- little girl climbing the stairs to her playhouse 4- little girl climbing into wooden playhouse 5- girl going into wooden building	**(XceptionNet+BiLSTM)** young girl in white shirt and gray pants and cane in chair and white pants on chair
	(ResNet+BiLSTM) little girl climbing into wooden cabin
	(Inception+BiLSTM) little girl climbing into wooden playhouse
	(EfficientNet+BiLSTM) little girl climbing into wooden playhouse
 1- small girl in the grass plays with fingerpaints in front of white canvas with rainbow on it. 2- little girl covered in paint sits in front of painted rainbow with her hands in bowl 3- there is girl with pigtails sitting in front of rainbow painting 4- little girl is sitting in front of large painted rainbow 5- young girl with pigtails painting outside in the grass	**(XceptionNet+LSTM)** little girl in pigtails plays on large sidewalk
	(ResNet+BiLSTM) young girl is sitting in front of large painted rainbow
	(Inception+BiLSTM) little girl with pigtails painting in the grass
	(EfficientNet+BiLSTM) little girl is sitting in front of large painted rainbow
	(ResNet+BiLSTM) man stands in front of very tall building.

<table>
<tr><th>Image & Original Description</th><th>Predicted Description</th></tr>
<tr><td>
1- there is skyscraper in the distance with man walking in front of the camera
2- behind the man in red shirt stands large skyscraper
3- man stands in front of very tall building
4- man is standing in front of skyscraper
5- man stands in front of skyscraper</td><td>(XceptionNet+LSTM)
man in white suit stands in the sidewalk.
(Inception+BiLSTM)
man in red shirt stands on the camera
(EfficientNet+BiLSTM)
man stands in front of skyscraper</td></tr>
<tr><td>
1- three dogs are standing in the grass and person is sitting next to them.
2- three dogs stand in grassy field while person kneels nearby.
3- three dogs are playing on grassy hill with blue sky
4- woman crouches near three dogs in field
5- three dogs on grassy hill</td><td>(XceptionNet+LSTM)
three dogs are standing on grassy grassy grassy grassy
(ResNet+BiLSTM)
three dogs are playing in the grass and kneels nearby
(Inception+BiLSTM)
three dogs are playing in grassy field
(EfficientNet+BiLSTM)
three dogs on grassy hill</td></tr>
</table>

Image & Original Description	**Predicted Description**
 1- boy in blue swimming trunks slides down yellow slide into wading pool with inflatable toys floating in the water 2- child is falling off slide onto colored balloons floating on pool of water 3- boy sliding down slide into pool with colorful tubes 4- boy in blue shorts slides down slide into pool 5- boy rides down slide into small backyard pool	**(Xception+LSTM)** boys in swimming pool **(ResNet+BiLSTM)** boys in swimming pool of water **(Inception+BiLSTM)** boys in swimming pool with inflatable toys **(EfficientNet+BiLSTM)** boys in swimming pool

Figure 4.5. Results of testing the image captioning MobileNet-LSTM, MobileNet-GRU, and MobileNet-GRU Feature Fusion models using some of test set samples.

All these new models achieved a lower performance compared to the VGG-16 and MobileNet models in terms of all performance metrics. Table 4.4 concludes the performance evaluation results using BLEU, ROUGE, CIDEr, METEOR, and Loss of many image captioning models.

Table 4.4. Results of training the many image captioning models.

Model	**Loss**	**Val-Loss**	**BLEU-1**	**BLEU-2**	**ROUGE**	**CIDEr**	**METEOR**
XceptionNet + LSTM	5.7445	4.6259	0.378521	0.183920	0.2876	0.1015	0.2867
ResNet50 + LSTM	5.1031	5.8784	0.426991	0.195069	0.2696	0.0899	0.2712

IneptionV3 + LSTM	4.7992	5.8002	0.477224	0.477224	0.2752	0.1001	0.2829
EfficientNet + LSTM	0.5293	8.7210	0.474467	0.226559	0.2759	0.1080	0.2931

4.6. RESULTS OF MODIFICATIONS ON THE BEST MODEL VGG-GRU FEATURE FUSION

In this section, the modifications on the best image captioning model will be applied, and the performance will be evaluated.

The proposed modifications are:

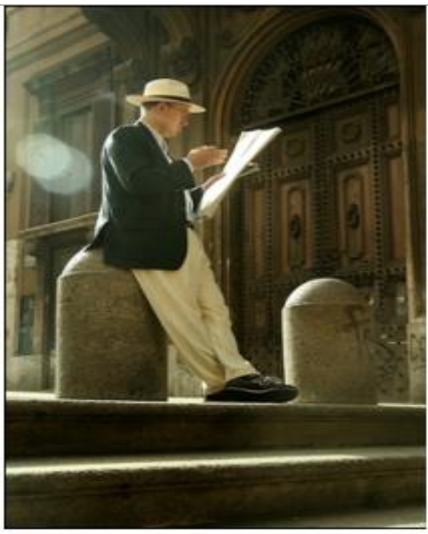
- Minimizing the dictionary size to only 15000 words by getting rid of phrases that are unlikely to be relevant; over-fitting can be avoided.
- Adding an attention layer to the language model to allow the model to focus on different parts of the image when generating the captions, which can improve the performance of the image captioning process. Table 4.5. concludes the performance evaluation results using BLEU, ROUGE, CIDEr, METEOR, and Loss of many image captioning models.

Table 4.5. Results of performance evaluation using BLEU, OUGE, CIDEr, METEOR, and Loss of many captioning systems.

Model	Loss	Val-Loss	BLEU-1	BLEU-2	ROUGE	CIDEr	METEOR
VGG16 + GRU model with feature fusion reduced vocabulary	4.4177	4.1833	0.663963	0.401557	0.3150	0.2383	0.3453

VGG16 + GRU Attention based model with feature fusion reduced vocabulary	4.3729	4.04287	0.673577	0.371224	0.3353	0.2224	0.3377

Figure 4.6. shows some examples of image captioning using some test samples with a comparison of the best two models (VGG+GRU with feature fusion and VGG+GRU with attention layer and filtered vocabulary).

Model	VGG16 + GRU model with feature fusion reduced vocabulary	VGG16 + GRU Attention based model with feature fusion reduced vocabulary
 1- nicely dressed man reading the newspaper in front of an older building 2- guy leaning on structure in front of building reading something 3- man in dark jacket and hat reading paper while leaning 4- man with hat leaning against post reading newspaper 5- man in hat and black jacket reads newspaper	man in black shirt and jeans is standing on the street	man in blue shirt is sitting on bench

Model	**VGG16 + GRU model with feature fusion reduced vocabulary**	**VGG16 + GRU Attention based model with feature fusion reduced vocabulary**
 1- crowd of people looking towards person in gorilla costume inside building 2- group of people travel through line along side gorilla statue 3- crowd waits to interact with some people in an indoor location 4- there is group of people looking at life sized gorilla 5- group of people are lined up inside waiting	group of people are sitting at table	group of people are standing in front of the people

Model	**VGG16 + GRU model with feature fusion reduced vocabulary**	**VGG16 + GRU Attention based model with feature fusion reduced vocabulary**
 1- two men in black pants black buttondown shirts and ties stand in front of an unmarked brown doorway 2- two men dressed in black pants and shirts are lounging outside door 3- two men taking break at the back of business 4- two men standing outside next to building 5- two men taking break during work	man in blue shirt is sitting on the sidewalk	man in black shirt and blue shirt is standing on the street

Model	VGG16 + GRU model with feature fusion reduced vocabulary	VGG16 + GRU Attention based model with feature fusion reduced vocabulary
 1- group of people walk in relaxed line across brick paved courtyard 2- group of people wait in long line in park to view statue 3- line of people makes its way up the steps 4- the line of waiting people is very long 5- crowd of people are waiting in line	group of people are walking down the street	group of people are walking down the street

Model	VGG16 + GRU model with feature fusion reduced vocabulary	VGG16 + GRU Attention based model with feature fusion reduced vocabulary
 1- child in pink dress is climbing up set of stairs in an entry way 2- little girl in pink dress going into wooden cabin 3- little girl climbing the stairs to her playhouse 4- little girl climbing into wooden playhouse 5- girl going into wooden building	girl climbing into wooden cabin	girl climbing into wooden cabin

Model	**VGG16 + GRU model with feature fusion reduced vocabulary**	**VGG16 + GRU Attention based model with feature fusion reduced vocabulary**
 1- small girl in the grass plays with fingerpaints in front of white canvas with rainbow on it. 2- little girl covered in paint sits in front of painted rainbow with her hands in bowl 3- there is girl with pigtails sitting in front of rainbow painting 4- little girl is sitting in front of large painted rainbow 5- young girl with pigtails painting outside in the grass	little girl in paint sits in painted painted rainbow with her hands in bowl	little girl in paint sits in painted painted rainbow with her hands in bowl

Model	VGG16 + GRU model with feature fusion reduced vocabulary	VGG16 + GRU Attention based model with feature fusion reduced vocabulary
 1- there is skyscraper in the distance with man walking in front of the camera 2- behind the man in red shirt stands large skyscraper 3- man stands in front of very tall building 4- man is standing in front of skyscraper 5- man stands in front of skyscraper	man in blue shirt and jeans is walking through front of skyscraper	man is standing in front of skyscraper

Model	**VGG16 + GRU model with feature fusion reduced vocabulary**	**VGG16 + GRU Attention based model with feature fusion reduced vocabulary**
 1- three dogs are standing in the grass and person is sitting next to them. 2- three dogs stand in grassy field while person kneels nearby. 3- three dogs are playing on grassy hill with blue sky 4- woman crouches near three dogs in field 5- three dogs on grassy hill	three dogs are playing in grassy field with cow kneels in the background	two dogs are walking down the street

Figure 4.6. Comparison of VGG-16 with GRU and VGG-16 with Attention and filtered vocabulary.

Although the BLEU-1 score of the VGG-16 with attention model is higher than the best VGG model (VGG-16 with GRU), the BLEU-2 score is less, and this is normal as seen in Figure (4-6) where the description sentences of the test samples are more accurate in case of VGG-GRU than in the attention model.

4.7. DISCUSSION OF THE IMAGE CAPTIONING RESULTS

In order to make a good discussion of the proposed models and the corresponding results, each performance metric is discussed among all models.

The discussion will take into account the comparison between all trained image captioning models in all scenarios.

All models will be individually compared to each other in terms of all evaluation metrics (Validation loss, BLEU-1, BLEU-2, ROUGE, CIDEr, and METEOR), and the conclusion of the comparison will be discussed.

Table 4.6. shows this study's performance metrics for all proposed image captioning models.

Table 4.6. Performance metrics for all proposed image captioning models.

Model	**Loss**	**Val-Loss**	**BLEU-1**	**BLEU-2**	**ROUGE**	**CIDEr**	**METEOR**
VGG+LSTM	2.2245	6.1677	0.458519	0.203514	0.3049	0.2883	0.3162
VGG+BiLSTM	4.4897	4.8086	0.498595	0.255005	0.2995	0.1092	0.3061
VGG+BiLSTM Different training Parameter	4.4358	4.7641	0.523388	0.269940	0.3010	0.1158	0.3090
VGG+GRU	3.4589	4.19689	0.620133	0.387619	0.3278	0.2925	0.3697
VGG+GRU Feature Fusion	4.4177	4.1833	0.663963	0.401557	0.3150	0.2383	0.3453
MobileNet + BiLSTM	4.6787	6.1725	0.516679	0.244216	0.2987	0.1096	0.3053
MobileNet + GRU	4.1251	4.1251	0.654240	0.352221	0.3283	0.3345	0.3593

Model	Loss	Val-Loss	BLEU-1	BLEU-2	ROUGE	CIDEr	METEOR
MobileNet + GRU Feature Fusion	4.2774	4.1764	0.649498	0.380580	0.3185	0.2044	0.3313
XceptionNet + LSTM	5.7445	4.6259	0.378521	0.183920	0.2876	0.1015	0.2867
ResNet50 + LSTM	5.1031	5.8784	0.426991	0.195069	0.2696	0.0899	0.2712
IneptionV3 + LSTM	4.7992	5.8002	0.477224	0.228147	0.2752	0.1001	0.2829
EfficientNet + LSTM	0.5293	8.7210	0.474467	0.226559	0.2759	0.1080	0.2931
VGG16 + GRU Attention based model with feature fusion reduced vocabulary	4.3729	4.04287	0.673577	0.371224	0.3353	0.2224	0.3377

To discuss the results, each performance metric will be taken into account individually, and a judgment of all image captioning models based on each metric will be made.

Figure 4.7 includes the comparison of the validation loss (Val-Loss) metric, in which the VGG16+GRU with attention layer achieved the lowest validation loss with 4.04287 value while the worst model was the EfficeintNet+LSTM model with 8.7210.

Some other models, like VGG+GRU and MobileNet+GRU, also achieved small validation loss values, which means that these models are better at describing images since the validation loss describes the error value of the validation set during the training process.

However, validation loss cannot reflect the entire truth about the image captioning models; instead, other metrics should be examined, like BLEU.

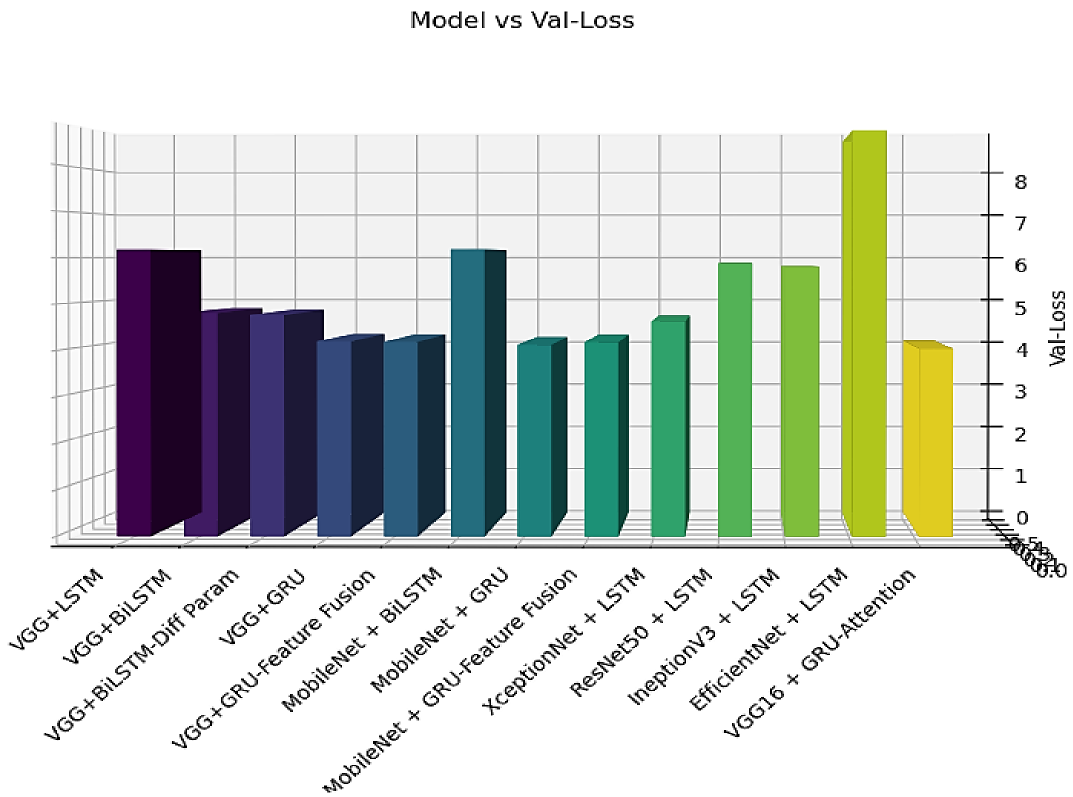


Figure 4.7. Comparison of all proposed image captioning models using Validation Loss.

Now, in terms of BLEU-1 performance metric, Figure 4.8. shows that the best model with the highest BLEU-1 score is the VGG+GRU with Attention, and the second best model is the VGG+GRU with feature fusion.

Other good models like MobileNet+GRU and MobileNet+GRU with feature fusion also achieved a closed BLEU-1 value to the attention model BLUE-1 score. Since the BLEU-1 score calculates the overlap of only a single word between the original description and the generated one, the results shown in Figure (4-8) mean that the models (VGG+GRU with Attention, MobileNet+GRU and MobileNet+GRU with feature fusion) have the best overlap of single word of the generated description against the original one.

To conclude:

These models are the best models that can capture the individual words that already exist in the original captions.

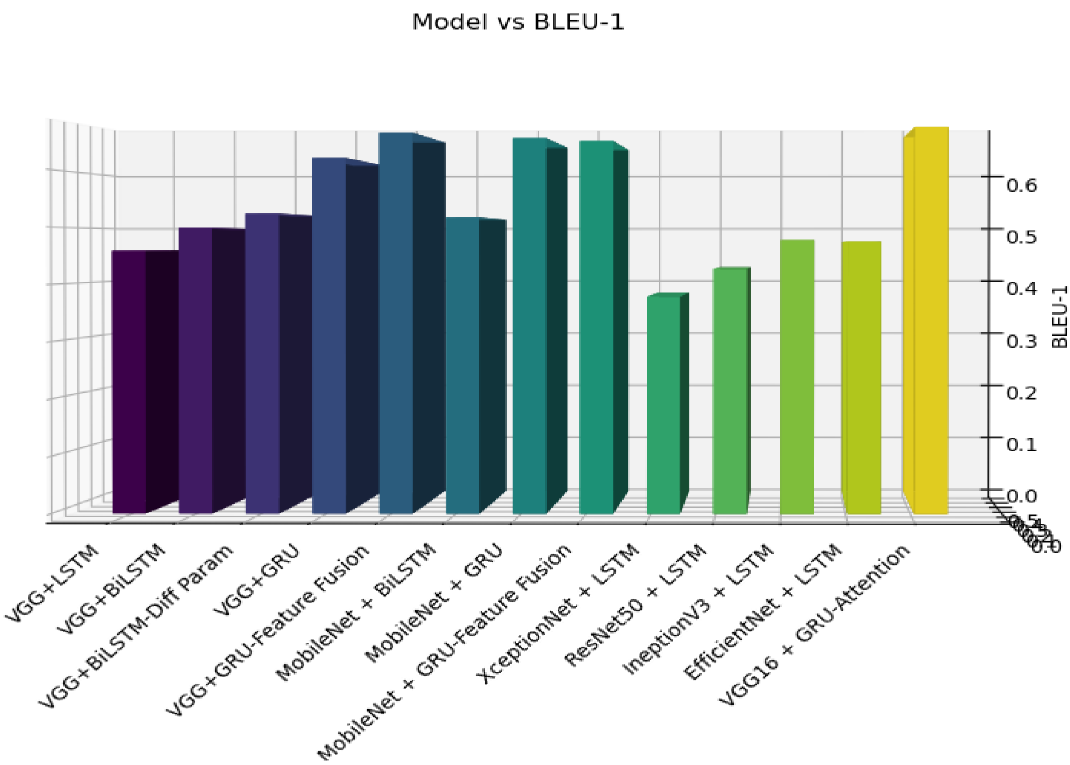


Figure 4.8. Comparison of all proposed image captioning models using BLEU-1.

For the BLEU-2 metric (Figure (4-9)), the calculations almost led to the same conclusion of the BLEU-1 score except for the VGG16+GRU with attention model in which the BLEU-2 score is less than other models like VGG+GRU with feature fusion, VGG+GU and MobileNet+ GRU with Feature fusion. However, this exception means that although the BLEU-1 score of the VGG+GRU with attention model achieves the best BLEU-1 score, its ability to capture the coherence between words of the describing caption are lower than other models.

The same remark is noticed for the VGG-GRU with Feature Fusion, in which the BLEU-1 score was less than VGG-GRU with Attention. However, in terms of the

BLEU-2 score, this model achieved the best result, indicating the captioning sentences' powerful grammatical and contextual coherence.

This means this model's words reflect the same coherence and relationships in the original sentence.

Let's check this predicted description: " little girl in paint sits in painted rainbow with her hands in bowl" while the original sentence is " little girl covered in paint sits in front of painted rainbow with her hands in bowl." These two descriptions are too closed but do not use exactly the same words; for example, "Covered" and "front" are not mentioned in the predicted sentence. However, the semantics of the sentences are exactly the same.

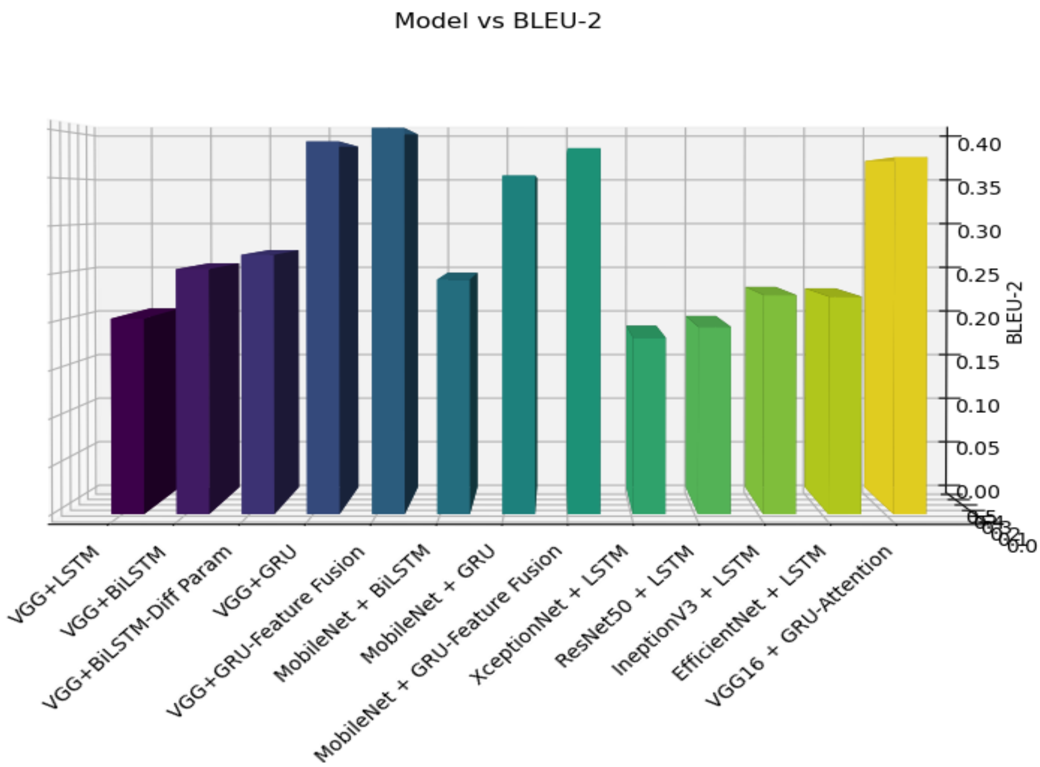


Figure 4.9. Comparison of all proposed image captioning models using BLEU-2.

For the third performance measure, "ROUGE" which is described in Figure 4.10. the best model with the highest ROUGE is the VGG16+GRU with attention model, which also achieved the best performance in terms of BLEU-1.

Other models like VGG+GRU and MobileNet+GRU also achieved high ROUGE values.

These results indicate that these models contain more words of the original description sentences in their prediction captions. More words in the original sentence mean a better captioning system.

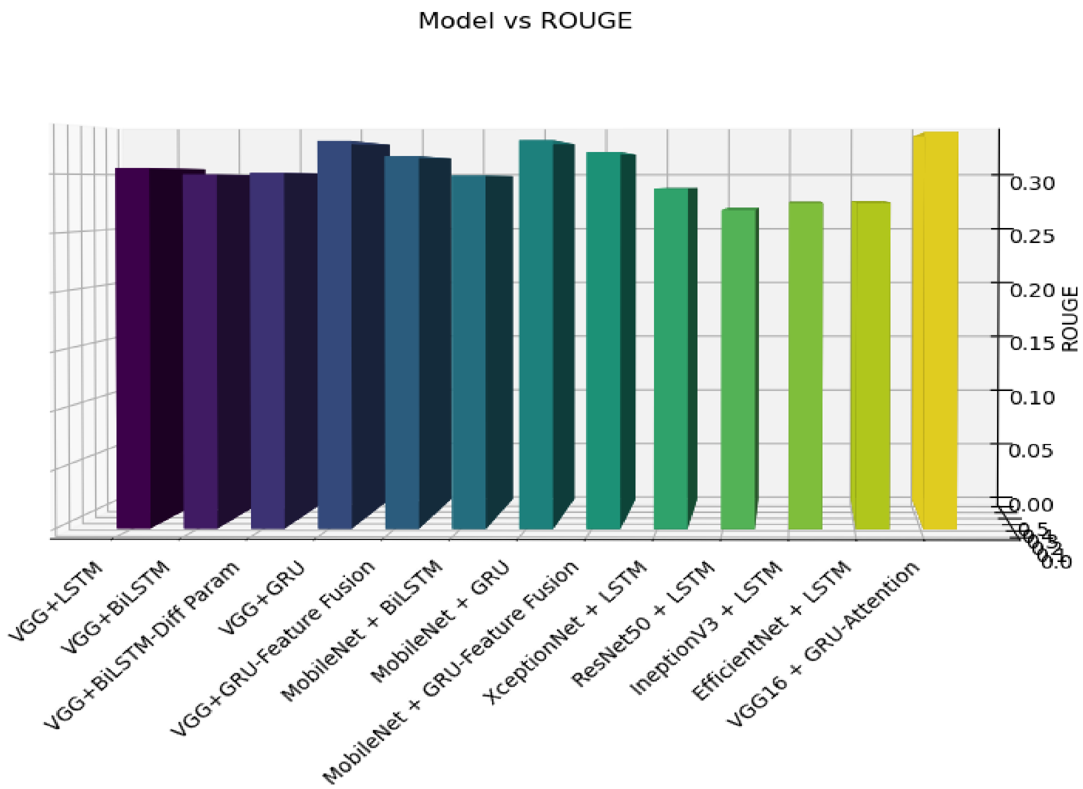


Figure 4.10. Comparison of all proposed image captioning models using ROUGE

According to the results, the CIDEr value of the MobileNet + GRU model was the best one, as illustrated in Figure (4-11). However, other models like VGG+GRU and VGG+LSTM also achieved a good CIDEr value, meaning that these models produce captioning sentences very close to human descriptions. Besides that, these models are more accurate at generating semantic words in comparison with the original words of the description sentences.

Let us check the results of the description of two models, one with high CIDEr and another with low CIDEr. In the image with the original description, " man stands in front of skyscraper", we got the following results:

Man stands in front of skyscraper (In case of MobeileNet+GRU)
Man in white suit stands in the sidewalk (In case of XceptionNet+LSTM)

The difference between the two description results is obvious since the XceptionNet+LSTM provides a very far description of the man in the scene. However, some words in the description of XceptionNet exist in the original description, like "man", "stands," "in" but the description of the situation of the man is totally wrong. CIDEr in the case of XceptionNet, is lower than MobileNet+GRU.

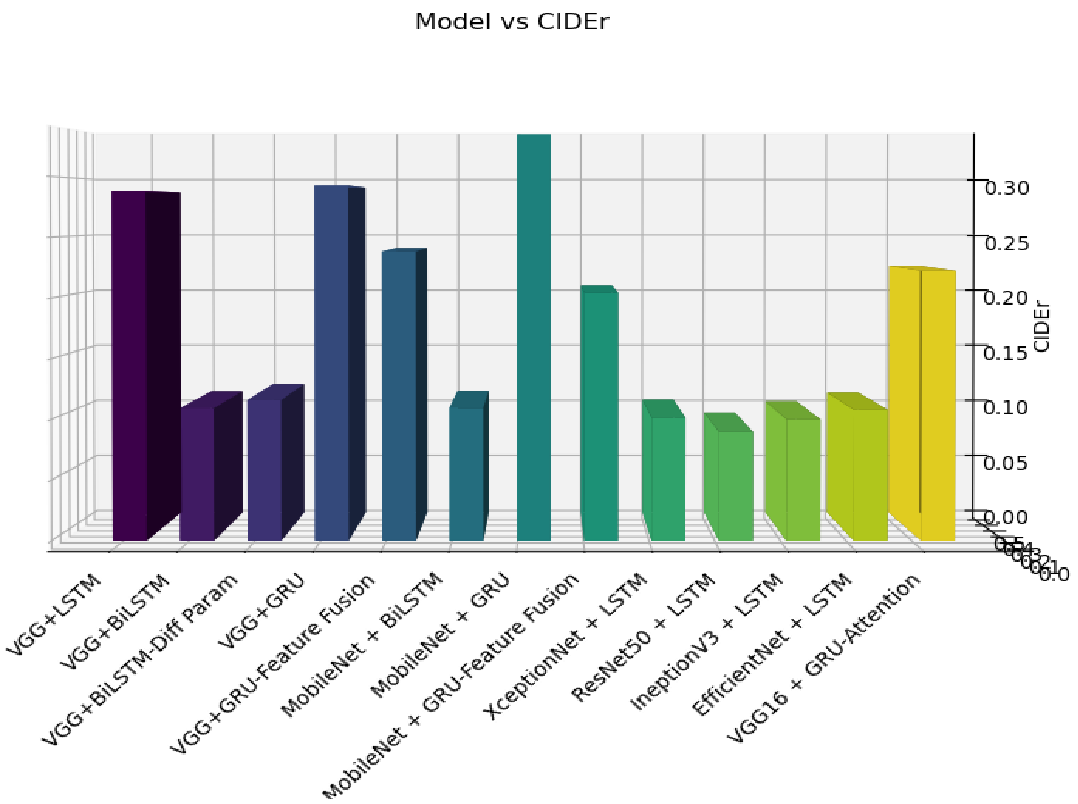


Figure 4.11. Comparison of all proposed image captioning models using CIDEr

The final performance metric is the METEOR (Figure (4.12)) in which the VGG+GRU model achieved the best score. Other models like MobileNet+GRU, VGG+GRU with Feature Fusion, and VGG16+GRU with Attention also achieved good MOTEOR values, indicating their ability to generate captions with high precision and recall values, meaning that these models are good at capturing linguistic variations and nuances.

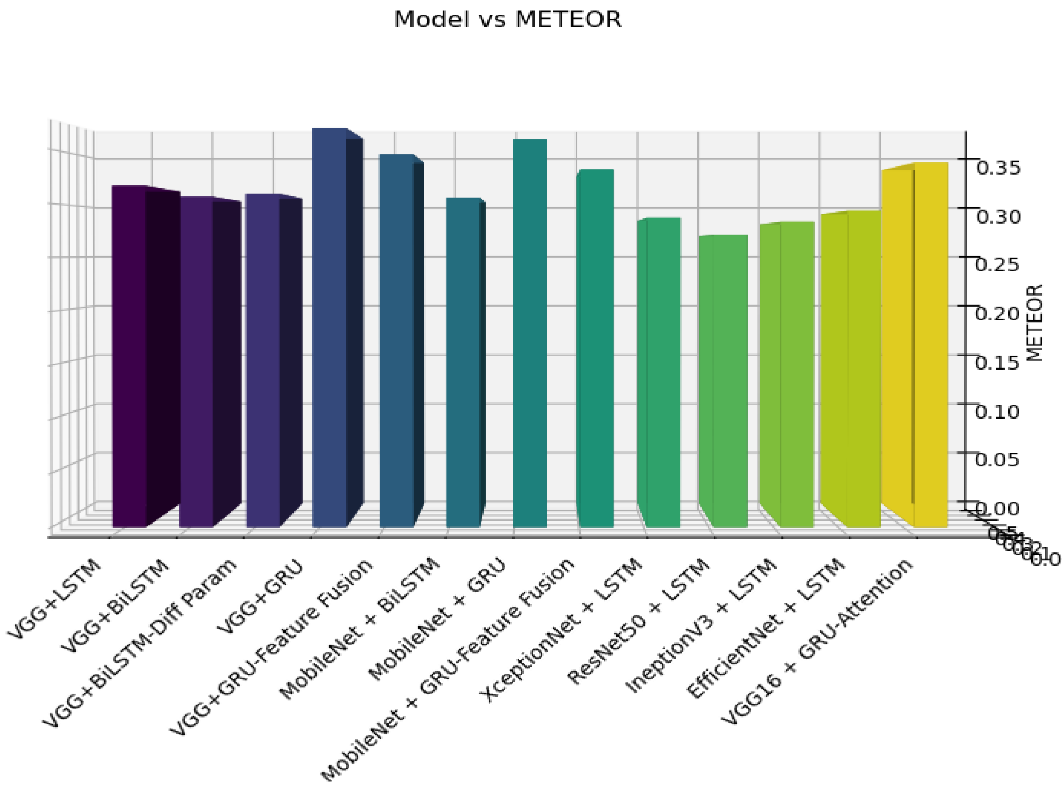


Figure 4.12. Comparison of all proposed image captioning models using METEOR.

4.8. COMPARISON WITH RELATED STATE-OF-THE-ART

Table 4.7. Stands for a general comparison between the current study and previous related work in the field of image captioning and description.

Table 4.7. General comparison between the current study and related state-of-art.

Study	Methodologies	Dataset	Results	Notes or limitations
Sammani et .al 2020 [31]	EditNet, copy-LSTM with Attention	MSCOCO dataset with a total of 82783 training images, 40504	BLEU-1: 77.9 BLEU-2: 38 CIDEr: 0.012 (1.2) SPICE: 21.2	time-consuming due to the refinement process

		validation and 40775 test images		
Khan et al. 2021 [32]	ResNet-50 and 1D-CNN	BanglaLekha Image Captions dataset, which consisted of 9000 images	BLEU-1: 0.651 (65.1) CIDEr: 0.572 METEOR: 0.297 ROUGE: 0.43	Recognize only humans in the scene
AK Poddar et al. 2023 [33]	VGG16 + LSTM	Flickr8k Hindi dataset with 8000 training and 100 validation images	BLEU-1: 0.359 (35.9) to 0.55 (55)	The used dataset has a moderate size. The model was evaluated using only one metric, "BLEU"
Wang et .al 2022 [35]	InceptionV3 and Bi-LSTM	Flickr30k	BLEU-1: 65.9	No combine of the visual and text attention mechanisms in their model
Selivanov et al.2023 [37]	GPT3	MSCOCO	BLEU-1: 0.725 (72.5)	No state-of-art comparison between their study and others on the same medical datasets
Xie et al. 2023 [39]	Bi-LSTM Attention Fast region CNN	Flickr30k	BLEU-1: 64.5	They compared Bi-LSTM with

				Bi-LSTM with Attention
Current Study	VGG-16, MobileNet, InceptionV3, ResNet50, Xception, For language models: LSTM, Bi-LSTM, GRU, GRU attention, GRU with Feature fusion of image and captions features	Flickr30k	Best BLEU-1: 0.674 (67.4) BLEU-2: 0.402 (40.2) ROUGE: 0.3353 CDIEr: 0.3345 METEOR: 0.3453	Based on lightweight image and language model. Used one dataset.

Most previous studies utilized on the Flickr30k dataset have many limitations in computational time or accuracy. Most of the previous studies did not consider all performance metrics used in the image captioning process to make a comprehensive judgment. Some of the previous studies utilized different datasets like MSCOCO and other specific datasets and achieved different performances.

The current study outperforms most previous studies, especially those on the same dataset, Flickr30K. The current study made a comprehensive comparison of many lightweight image captioning models and defined the best one.

The current study utilized the idea of feature fusion of image and caption features in one feature vector to minimize training time and improve performance. The current study also investigates the efficiency of using attention layers with the GRU language model under a filtered vocabulary to avoid overfitting and improve performance.

PART 5

CONCLUSION AND FUTURE WORK

In this study, comprehensive image captioning and description models were built, trained, and evaluated using many image captioning metrics. Initially, the Flickr30K dataset was acquired. Subsequently, preprocessing was applied to both images and captions. This involved resizing images and splitting, cleaning padding, and encoding captions. In specific scenarios, the captions were filtered to get the most frequent 15000 words out of all vocabulary words. In the next step, the image model was built using many pre-trained models (VGG-16, MobileNet, InceptionV3, XceptionNet, ResNet50).

The image features were extracted using these models after eliminating the final classification layers. These extracted features were then fed into the next step. For the language model, various options were utilized, such as LSTM, Bi-LSTM, GRU, and GRU, with attention layers. In some experiments, image and caption features were fused to create a unified feature vector. Other scenarios involved concatenation (vector after vector) of the image and caption features. In some scenarios, the entire vocabulary was used, while in others, the filtered vocabulary was used. Experiments utilized a training set comprising 80% of the Flickr30K dataset, with the remaining 20% used as a test set. Results proved the high performance of several proposed models, including VGG-GRU, VGG-GRU with feature fusion, VGG-GRU with attention and filtered vocabulary, and MobileNet-GRU models. The highest registered BLEU-1 score corresponded to the VGG-GRU with attention model with a 0.674 value, while the best BLEU-2 score was achieved by the VGG+GRU Feature Fusion with a 0.402 value. Other metrics like ROUGE, CIDEr, and METEOR were also been used to compare the models together in terms of many captioning concepts. The current study was also compared with related state-of-the-art studies. This comparison proved the efficiency and high performance of the study.

Future studies can benefit from the limitations of the current study. Here are some of the recommendations for future work:

- Using a fusion of lightweight and heavyweight image and language models to achieve both good accuracy and moderate computational time.
- Try different image captioning datasets.
- Work deeper with the attention model by developing a new language model with the benefit of attention models.

REFERENCES

[1] Das, B., Pal, R., Majumder, M., Phadikar, S., and Sekh, A. A., "A visual attention-based model for bengali image captioning", SN Computer Science, 4 (2): 208 (2023).

[2] Rinaldi, A. M., Russo, C., and Tommasino, C., "Automatic image captioning combining natural language processing and deep neural networks", ***Results In Engineering***, 18: 101107 (2023).

[3] Zouitni, C., Sabri, M. A., and Aarab, A., "A Comparison Between LSTM and Transformers for Image Captioning " in ***International Conference on Digital Technologies and Applications***,: Springer, pp. 492-500 , (2023).

[4] Degadwala, S., Vyas, D., Biswas, H., Chakraborty, U., and Saha, S., "Image captioning using inception V3 transfer learning model" *6th* ***International Conference on Communication and Electronics Systems (ICCES)***, IEEE, pp. 1103-1108. (2021).

[5] Marzouk, R., Alabdulkreem, E., Nour, M. K., Al Duhayyim, M., Othman, M., Zamani, A. S., Yaseen, I., and Motwakel, A., "Natural Language Processing with Optimal Deep Learning-Enabled Intelligent Image Captioning System", ***Computers, Materials & Continua***, 74 (2): (2023).

[6] Hilal, A. M., Alrowais, F., Al-Wesabi, F. N., and Marzouk, R., "Red Deer Optimization with Artificial Intelligence Enabled Image Captioning System for Visually Impaired People", ***Computer Systems Science & Engineering***, 46 (2): (2023).

[7] Sharma, A., Chaudhary, A., and Dixit, A., "Image Captioning Using Python", ***International Conference on Power, Instrumentation, Energy and Control (PIECON),*** IEEE, pp. 1-5., (2023).

[8] Xu, L., Tang, Q., Lv, J., Zheng, B., Zeng, X., and Li, W., "Deep Image Captioning: A Review of Methods, Trends and Future Challenges", ***Neurocomputing***, 126287 (2023).

[9] He K., Z. X. and Ren S., & S. J., "Deep Residual Learning for Image Recognition", ***Proceedings of the IEEE conference on computer vision and pattern recognition.*** (2016).

[10] Simonyan, K. and Zisserman, A., "Very deep convolutional networks for large-scale image recognition", ***ArXiv Preprint ArXiv:1409.1556***, (2014).

[11] Tan, Mingxing, and Quoc Le. "Efficientnet: Rethinking model scaling for convolutional neural networks." In International conference on machine learning, pp. 6105-6114. PMLR, (2019).

[12] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y., "Generative adversarial nets", ***Advances In Neural Information Processing Systems***, 27: (2014).

[13] Z. A. khalaf, and N. T. A. Ramaha, "Review Of Breast Diagnosis Detection and Classification Based on Machine Learning," International Conference on Trends in Advanced Research, vol. 1, pp. 222–230, Mar. 2023. [Online]. Available: https://as-proceeding.com/index.php/ictar/article/view/209.

[14] Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., and Adam, H., "Mobilenets: Efficient convolutional neural networks for mobile vision applications", ***ArXiv Preprint ArXiv:1704.04861***, (2017).

[15] Ramaha, N. T. A., Mahmood, R. M., Hameed, A. A., Fitriyani, N. L., Alfian, G., and Syafrudin, M., "Brain Pathology Classification of MR Images Using Machine Learning Techniques", ***Computers***, 12 (8): 167 (2023).

[16] Thangavel, K., Palanisamy, N., Muthusamy, S., Mishra, O. P., Sundararajan, S. C. M., Panchal, H., Loganathan, A. K., and Ramamoorthi, P., "A novel method for image captioning using multimodal feature fusion employing mask RNN and LSTM models", ***Soft Computing***, 1–14 (2023).

[17] Dubey, S., Olimov, F., Rafique, M. A., Kim, J., and Jeon, M., "Label-attention transformer with geometrically coherent objects for image captioning", ***Information Sciences***, 623: 812–831 (2023).

[18] Gan, Zhe, Chuang Gan, Xiaodong He, Yunchen Pu, Kenneth Tran, Jianfeng Gao, Lawrence Carin, and Li Deng. "Semantic compositional networks for visual captioning." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 5630-5639. (2017).

[19] Liao, W., Hu, K., Yang, M. Y., and Rosenhahn, B., Liao, "Text to image generation with semantic-spatial aware gan." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. (2022).

[20] Yusuf, A. A., Chong, F., and Xianling, M., "An analysis of graph convolutional networks and recent datasets for visual question answering", ***Artificial Intelligence Review***, 55 (8): 6277–6300 (2022).

[21] Mozes, M., Schmitt, M., Golkov, V., Schütze, H., and Cremers, D., "Scene Graph Generation for Better Image Captioning?", ***ArXiv Preprint ArXiv:2109.11398***, (2021).

[22] Zhang, Y., Pan, Y., Yao, T., Huang, R., Mei, T., and Chen, C.-W., "Boosting scene graph generation with visual relation saliency", ***ACM Transactions On Multimedia Computing, Communications And Applications***, 19 (1): 1–17 (2023).

[23] Suresh, K. R., Jarapala, A., and Sudeep, P. V, "Image Captioning Encoder–Decoder Models Using CNN-RNN Architectures: A Comparative Study", ***Circuits, Systems, And Signal Processing***, 41 (10): 5719–5742 (2022).

[24] Kalra, S. and Leekha, A., "Survey of convolutional neural networks for image captioning", ***Journal Of Information And Optimization Sciences***, 41 (1): 239–260 (2020).

[25] Wang, H., Wang, H., and Xu, K., "Evolutionary recurrent neural network for image captioning", ***Neurocomputing***, 401: 249–256 (2020).

[26] Staudemeyer, R. C. and Morris, E. R., "Understanding LSTM--a tutorial into long short-term memory recurrent neural networks", ***ArXiv Preprint ArXiv:1909.09586***, (2019).

[27] N. Gupta and A. S. Jalal, "Integration of textual cues for fine-grained image captioning using deep CNN and LSTM," ***Neural Computing and Applications,*** vol. 32, pp. 17899-17908, (2020).

[28] Sharma, H., Agrahari, M., Singh, S. K., Firoj, M., and Mishra, R. K.,"Image captioning: a comprehensive survey."International Conference on Power Electronics & IoT Applications in Renewable Energy and its Control (PARC). IEEE, (2020).

[29] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I., "Attention is all you need", ***Advances In Neural Information Processing Systems***, 30: (2017).

[30] Liu, W., Chen, S., Guo, L., Zhu, X., and Liu, J., "Cptr: Full transformer network for image captioning", ***ArXiv Preprint ArXiv:2101.10804***, (2021).

[31] Sammani, F. and Melas-Kyriazi, L., "Show, edit and tell: a framework for editing image captions." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition , (2020).

[32] Khan, F., Mohammad, S. M. S.-U.-R., and Islam, M. S., "Improved bengali image captioning via deep convolutional neural network based encoder-decoder model," presented at the ***International Joint Conference on Advances in Computational Intelligence***: IJCACI Singapore, (2021).

[33] Poddar, A. K. and Rani, R., "Hybrid Architecture using CNN and LSTM for Image Captioning in Hindi Language", ***Procedia Computer Science***, 218: 686–696 (2023).

[34] He, S., Liao, W., Tavakoli, H. R., Yang, M., Rosenhahn, B., and Pugeault, N., "Image captioning through image transformer" in ***Proceedings of the Asian conference on computer vision***, (2020).

[35] Wang, Z., Shi, S., Zhai, Z., Wu, Y., and Yang, R., "ArCo: Attention-reinforced transformer with contrastive learning for image captioning", ***Image And Vision Computing***, 128: 104570 (2022).

[36] Hu, J., Yang, Y., An, Y., and Yao, L., "Dual-Spatial Normalized Transformer for image captioning", ***Engineering Applications Of Artificial Intelligence***, 123: 106384 (2023).

[37] Selivanov, A., Rogov, O. Y., Chesakov, D., Shelmanov, A., Fedulova, I., and Dylov, D. V, "Medical image captioning via generative pretrained transformers", ***Scientific Reports***, 13 (1): 4171 (2023).

[38] Fei, Z., " Fei, "Attention-aligned transformer for image captioning." proceedings of the AAAI Conference on Artificial Intelligence . Vol. 36. No. 1., (2022).

[39] Xie, T., Ding, W., Zhang, J., Wan, X., and Wang, J., "Bi-LS-AttM: A Bidirectional LSTM and Attention Mechanism Model for Improving Image Captioning", ***Applied Sciences***, 13 (13): 7916 (2023).

[40] Patwari, N. and Naik, D.,"En-de-cap: An encoder decoder model for image captioning." 2021 5th International Conference on Computing Methodologies and Communication (ICCMC). IEEE., (2021).

[41] Mishra, S. K., Sinha, S., Saha, S., and Bhattacharyya, P., "Dynamic Convolution-based Encoder-Decoder Framework for Image Captioning in Hindi", ***ACM Transactions On Asian And Low-Resource Language Information Processing***, 22 (4): 1–18 (2023).

[42] Wang, Y., Xu, J., and Sun, Y.,"End-to-end transformer based model for image captioning." Proceedings of the AAAI Conference on Artificial Intelligence . Vol. 36. No. 3. (2022).

[43] Castro, R., Pineda, I., Lim, W., and Morocho-Cayamcela, M. E., "Deep learning approaches based on transformer architectures for image captioning tasks", ***IEEE Access***, 10: 33679–33694 (2022).

[44] Parvin, H., Naghsh-Nilchi, A. R., and Mohammadi, H. M., "Transformer-based local-global guidance for image captioning", ***Expert Systems With Applications***, 223: 119774 (2023).

[45] Sharma, D., Dhiman, C., and Kumar, D., "XGL-T transformer model for intelligent image captioning", ***Multimedia Tools And Applications***, 1–22 (2023).

[46] Yang, X., Wang, Y., Chen, H., Li, J., and Huang, T., "Context-Aware Transformer for image captioning", ***Neurocomputing***, 126440 (2023).

[47] Chen, C., Mu, S., Xiao, W., Ye, Z., Wu, L., and Ju, Q., "Improving image captioning with conditional generative adversarial nets." Proceedings of the AAAI conference on artificial intelligence . Vol. 33. No. 01., (2019).
[48] Amirian, S., Rasheed, K., Taha, T. R., and Arabnia, H. R.,"Image captioning with generative adversarial network." 2019 International Conference on Computational Science and Computational Intelligence (CSCI) . IEEE., (2019).
[49] Deepak, G., Gali, S., Sonker, A., Jos, B. C., Daya Sagar, K. V, and Singh, C., "Automatic image captioning system using a deep learning approach", ***Soft Computing***, 1–9 (2023).
[50] Honda, U., Watanabe, T., and Matsumoto, Y., "Switching to Discriminative Image Captioning by Relieving a Bottleneck of Reinforcement Learning." Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision ., (2023).
[51] Yan, S., Xie, Y., Wu, F., Smith, J. S., Lu, W., and Zhang, B., "Image captioning via hierarchical attention mechanism and policy gradient optimization", ***Signal Processing***, 167: 107329 (2020).
[52] Chen, T., Li, Z., Wu, J., Ma, H., and Su, B., "Improving image captioning with Pyramid Attention and SC-GAN", ***Image And Vision Computing***, 117: 104340 (2022).
[53] Padate, R., Jain, A., Kalla, M., and Sharma, A., "Image caption generation using a dual attention mechanism", ***Engineering Applications Of Artificial Intelligence***, 123: 106112 (2023).
[54] Babavalian, M. R. and Kiani, K., "Learning distribution of video captions using conditional GAN", ***Multimedia Tools And Applications***, 1–23 (2023).
[55] M. Alhamidi, and N. T. A. Ramaha, "Unveiling Alzheimer's Disease via MRI: Deep Learning Approaches for Accurate Detection," ***International Journal of Advanced Natural Sciences and Engineering Researches***, vol. 7, pp. 418-422, Oct. 2023.

[56] Young, P., Lai, A., Hodosh, M., and Hockenmaier, J., "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 (2014).

[57] Jia, J., Ding, X., Pang, S., Gao, X., Xin, X., Hu, R., and Nie, J., "Image captioning based on scene graphs: A survey", ***Expert Systems With Applications***, 120698 (2023).

[58] Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J.,"Bleu: a method for automatic evaluation of machine translation." Proceedings of the 40th annual meeting of the Association for Computational Linguistics ., (2002).

[59] Liu, S., Zhu, Z., Ye, N., Guadarrama, S., and Murphy, K.,"Improved image captioning via policy gradient optimization of spider." Proceedings of the IEEE international conference on computer vision ., (2017).

[60] Banerjee, S. and Lavie, A.,"METEOR: An automatic metric for MT evaluation with improved correlation with human judgments." Proceedings of the acl

workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization., (2005).

[61] Anderson, P., Fernando, B., Johnson, M., and Gould, S.,"Spice: Semantic propositional image caption evaluation." Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14. Springer International Publishing., (2016).

[62] M. Khalid, and N. T. A. Ramaha, "A Review on Image Captioning Using Deep Learning Methodologies: Limitations, Datasets, Metrics, and Recommendations,"***3rd GLOBAL CONFERENCE on ENGINEERING RESEARCH***, pp. 364-377, Sept. 2023.

RESUME

Zainab Khalid TAWFEEQ initiated her academic journey in Kirkuk, Iraq, and completed her secondary education at Barsh High School in the academic year 2015-2016. Subsequently, she pursued her undergraduate studies at Al-Iraqia University, graduating in the academic year 2019-2020. In 2021, she moved to Karabuk, Turkey, for her postgraduate studies and enrolled in the Master of Science program in Computer Engineering at Karabuk University.