symmetry

MDPI

Article

Generating Image Captions Using Bahdanau Attention Mechanism and Transfer Learning

Shahnawaz Ayoub [1,*], Yonis Gulzar [2,*], Faheem Ahmad Reegu [3] and Sherzod Turaev [4,*]

[1] Department of Computer Science and Engineering, Shri Venkateshwara University, NH-24, Venkateshwara Nagar, Gajraula 244236, Uttar Pradesh, India
[2] Department of Management Information Systems, College of Business Administration, King Faisal University, Al-Ahsa 31982, Saudi Arabia
[3] Department of Computer Science and Information Technology, Jazan University, Jazan 45142, Saudi Arabia
[4] Department of Computer Science & Software Engineering, College of Information Technology, United Arab Emirates University, Al Ain 15551, United Arab Emirates
* Correspondence: shahnawazayoub@outlook.com (S.A.); ygulzar@kfu.edu.sa (Y.G.); sherzod@uaeu.ac.ae (S.T.); Tel.: +966-545-719-118 (Y.G.)

Abstract: Automatic image caption prediction is a challenging task in natural language processing. Most of the researchers have used the convolutional neural network as an encoder and decoder. However, an accurate image caption prediction requires a model to understand the semantic relationship that exists between the various objects present in an image. The attention mechanism performs a linear combination of encoder and decoder states. It emphasizes the semantic information present in the caption with the visual information present in an image. In this paper, we incorporated the Bahdanau attention mechanism with two pre-trained convolutional neural networks—Vector Geometry Group and InceptionV3—to predict the captions of a given image. The two pre-trained models are used as encoders and the Recurrent neural network is used as a decoder. With the help of the attention mechanism, the two encoders are able to provide semantic context information to the decoder and achieve a bilingual evaluation understudy score of 62.5. Our main goal is to compare the performance of the two pre-trained models incorporated with the Bahdanau attention mechanism on the same dataset.

Keywords: image captioning; convolutional neural network; Bahdanau attention mechanism; natural language process

Citation: Ayoub, S.; Gulzar, Y.; Reegu, F.A.; Turaev, S. Generating Image Captions Using Bahdanau Attention Mechanism and Transfer Learning. *Symmetry* **2022**, *14*, 2681. https://doi.org/10.3390/sym14122681

Academic Editor: Changxin Gao

Received: 21 November 2022
Accepted: 12 December 2022
Published: 18 December 2022

1. Introduction

Image captioning or the automatic generation of descriptions of photographs in natural scenes is gaining attention in the domains of computer vision [1,2], natural language processing (NLP) [3,4], image indexing [5,6] and in supporting visually impaired people. NLP is a collective term relating to the automatic computational processing of human languages. It is described as the automatic interchange of natural language such as ordinary speech and text by software [7]. It is a difficult task that needs a thorough understanding of two different types of media data, namely vision and language [8]. Although captioning an image is easy for humans, it is very challenging for artificial intelligence (AI). Therefore, automatic captioning on a comprehensive understanding of real-world scenes of images is considered as a significant task. Semantic understanding of images by AI can help visually impaired persons with the brain–machine interface, autonomous/assisted driving, and intelligent navigation. As the NLP generates captions based on visual features retrieved in the deep learning network, the architecture of deep learning plays a critical role in captioning performance. Although neural networks have incredibly complicated architectures, deep learning methods provide an effective solution for data processing in these architectures.

Due to the wide range of applications, deep learning has been used in numerous areas such as healthcare, agriculture, etc. In healthcare, deep learning has been used for the prediction of disorders in children [9,10], disease classification [11,12] in the human body, and in agriculture, it is used for classification problems [13–15] or other image processing problems [16,17]. Deep learning architectures are also used in captioning to extract visual information from photos. They are then passed into NLP for caption generation. There are several studies for image captioning such as [18–20] that used Convolutional Neural Networks (CNN) and Long-Short Term Memory (LSTM). Most of the researchers have used the CNN-LSTM framework for image captioning. CNN acts as an image encoder interpreting visual areas and encoding them as area-specific features, whereas LSTM acts as a decoder and tries to understand all the generated words by the CNN. To capture the regions in an image, there are two ways [8]: One way is to divide an image based on the model's architecture, whereas the other way is an adaptive method in which a bounding box is used to capture the regions of an image at the object level.

However, it is very important to note down here that little visual information is required when forming non-visual words like "the", "were", and "itself", these attention models are nonetheless constrained to employ image attributes, which misleads the generation. Figure 1 shows the degree of importance of every word present in an image. The blue shade indicates the degree of importance. In the sentence "A girl is playing with her ball", more focus has to be given to the non-visual objects like "is", "with" and "her" as shown in Figure 1. The degree of importance of "is" in Figure 1 is 90%. Similarly, the degree of importance of "with" in the sentence is about 95%.

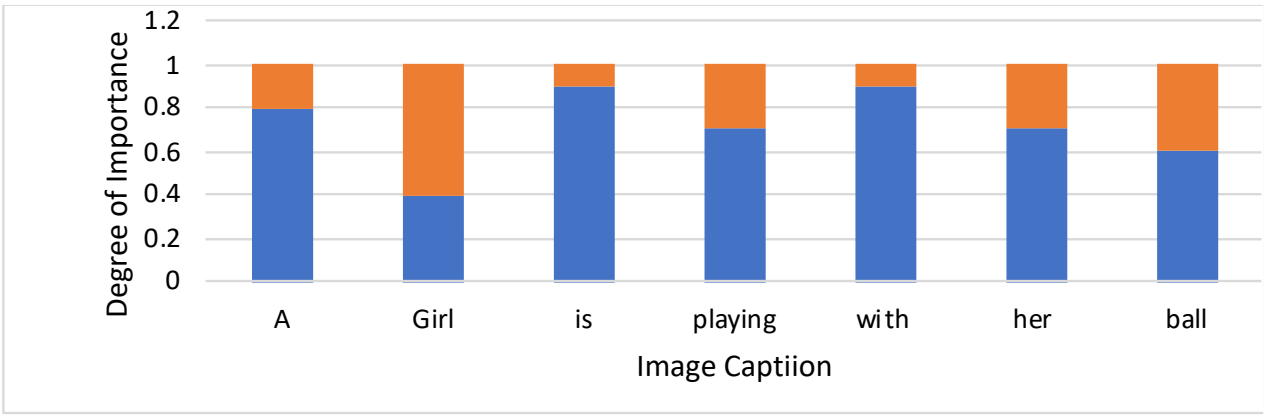


Figure 1. Degree of importance of non-visual objects like "is", "with", and "her".

It is worth noting here that the mechanism through which the features (visual as well as non-visual objects) of an image are utilized by a model is very crucial. While generating words for visual objects, the content of an image is to be focused, whereas while generating captions for non-visual words, the clues should be focused. The attention mechanism pays more attention to the non-visual feature vector as compared to the visual feature vector, therefore avoiding misleads. Thus, attention models are used to solve misleading problems and generate proper captions for non-visual objects present in an image. The literature reveals that an accurate semantic description of objects in all visual regions has attracted a lot of research interest, particularly in the field of attention mechanisms. For example, some works like [21] use feature extracted from previous words of non-visual words along with the visual words. As a result, the attention model should learn non-visual cues to aid the generation of non-visual words and ease the misleading problem [8].

Previous studies such as [18,20] used CNN for automatic captioning, but it is very important to reduce the computation and time. To improve the training and reduce the number of hyperparameters, a visual geometry group network (VGGNET) was proposed. The use of CNN with a smaller number of hyperparameters for caption generation has the

benefit of utilizing low hardware computation and training time. Therefore, on the one hand, we need to make sure that non-visual words are given equal importance, whereas on the other, we are curious to know how the architecture of a deep neural network affects the caption prediction performance.

To the best of our knowledge, this work is the first that compares the two different architectures of CNN using the Bahdanau attention mechanism for caption generation. Through our work, we show that using an attention model with CNN architecture has shown better results in terms of predicting the captions of images.

1.1. Motivation

As one of the most valuable visual resources, automatic image captioning by machine learning and deep learning models play an important role in many applications like remote sensing, social media platforms like Facebook could infer from pictures directly, mapping natural language to images, medical image understanding, creating dynamic websites, etc. In addition, owing to broad application of image captioning, a wide variety of objects are detected in images manually by remote satellites, and by doctors. It is very hard and extremely infeasible for experts to detect objects in an image. In order to accelerate the process of understanding images, the utmost use of machine learning and deep learning models is important. When given an image, an efficient image captioning system should be able to automatically detect visual and non-visual objects present and caption the image on that basis.

1.2. Related Work

Image captioning has captivated the interest of many researchers, and numerous models have been presented [22–24]. Due to significant developments in Deep Neural Networks (DNNs), most state-of-the-art techniques [25,26] use CNN as encoder and RNN as decoder, thus understanding images and generating captions. In order to extract high-level semantic features from an image, CNN as an encoder is used. To decode these extracted features and generate captions, RNN as a decoder is used. CNN-RNN has been widely employed in the field of computer vision as CNN is a standard technique for image processing. Only a few groups create their own networks from the ground up, while the majority rely on pre-existing architectures such as ResNet-50 or DenseNet-121. These pre-trained neural networks are frequently trained on huge, publicly available datasets, such as ImageNet, and can thus recognize features of an image. The use of pre-trained models may improve the accuracy and speed up the training time of a new model. This is because the models have already learned the important features of an image and are ready to be transferred to the new task without learning from scratch. VGG16 and InceptionV3 have been proven to be promising techniques for image captioning.

Generating captions is a dynamic process in which visual, as well as non-visual objects, are to be decoded with correct semantic information. Anderson et al. in [27] presented a new method called "bottom-up and top-down". The salient regions of an image are proposed by a bottom-up module, which is further represented using a convolutional feature vector. The top-down module consists of two LSTM networks. The first LSTM network is a top-down visual attention model, while the second LSTM network is a language model. In another work [28], the authors proposed a hybrid model based on LSTM and graph convolutional networks. To enrich image representation, the semantic and spatial relations have been merged. For this purpose, they have used two spatial and semantic graphs. Chen et al. [29] use structured language description to transform the problem of complex image retrieval into a dense captioning and scene graph. In this paper, a novel method for scene graph matching has been proposed, and a novel CBIR dataset for large-scale analysis is used. For image retrieval, a dense caption reasoning strategy with two stages is used. The first is dense caption generation, while the second is scene graph construction and reasoning. However, there are a few situations when the decoder has no knowledge of how to interpret non-visual words like "was", "here",

"put", etc. Therefore, in this case, the model still outputs a meaningless result. To overcome this problem, the attention mechanism plays a very significant role. Based upon human intuition, the attention mechanism has shown significant improvement in various tasks of sequence learning. It first computes a score for each candidate vector, then uses the Softmax function to normalize the scores to weights, and then applies these weights to the candidates to get the attention result, i.e., weighted average vector. Various attention mechanisms have been proposed like multi-level attention [30], like multi-head and self-attention and spatial and channel wise attention. Huang et al. [31] proposed a novel framework called Attention on Attention. They applied this mechanism to the encoder as well as the decoder. It helps in modeling better relationships between various objects in the encoder. From the decoder side, it helps to filter out insignificant results of attention. A hierarchical attention mechanism-based framework is proposed by Yan et al. in [32]. For optimization purposes, they employed Generative Adversarial Network (GAN) with a policy gradient algorithm. Khan et al. in [22] proposed a novel framework that extracts features from an image using a pre-trained CNN model. The image is converted into a feature vector. Further, for decoding images, the authors used Gated Recurrent Units (GRU). To allow learning to focus on a particular portion of an image, authors combined GRU with the Bahdanau attention mechanism. However, they have not discussed the concept of transfer learning. Table 1 shows the summary of the literature.

Table 1. Summarized literature survey.

S. No	Dataset	Year	Model	Attention Mechanism	Remarks
1.	ILSVRC-2012	2015	ConvNet [27]	×	No focus on non-visual objects
2.	ImageNet	2015	ResNet 101 [26]	×	No focus on attention mechanism
3.	Pascal VOC 2008, Flickr8k, Flickr30k, MSCOCO, SBU	2015	CNN-LSTM [33]	×	No focus on attention mechanism
4.	Pascal VOC 2012, Pascal Context, Pascal Pearson part, cityscape	2017	ResNet 101 [34]	×	Proposed model failed to capture boundaries
5.	COCO, Flickr30K, Flickr8k	2016	Bi-LSTM [19]	×	No focus on attention mechanism
6.	COCO, Flickr30K, Flickr8k	2017	CNN [35]	√	Proposed model requires more investigation to avoid overfitting
7.	COCO dataset Flickr30K, Flickr8k	2018	CNN-RNN [36]	√	Framework is not end to end
8.	MSCOCO, COCO stuff	2018	FCN-LSTM [8]	√	Did not evaluate different pre-trained models
9.	MSCOCO, Visual Genome dataset, VQA v2.0 dataset	2018	Faster R-CNN + ResNet 101-LSTM [21]	√	Proposed model is computationally costly
10.	COCO dataset	2018	Graph CNN-LSTM [28]	√	Evaluated their model only on one dataset
11.	MS-COCO, Flickr30K	2019	Reference LSTM [37]	×	Did not incorporate attention mechanism
12.	MS-COCO	2019	RCNN-LSTM [31]	√	Attention on attention may lead to loss of significant information
13.	MS-COCO	2019	CNN-LSTM [38]	√	Training GAN is computationally very costly
14.	Flickr30K, Flickr8k	2020	CNN-LSTM [18]	√	Did not evaluate different pre-trained models
15.	MS-COCO	2022	CNN-GRU [22]	√	Evaluated the model only on one dataset

An attention function is defined as a function that maps a query and a set of key-value pairs to an output. The output is generated as a weighted sum of the values, with the weight allocated to each value determined by the query's compatibility function with the relevant key. On the basis of the decoder's output, attention methods can be of three types:

(a) Attribute-based visual representation: The features of an image are represented by a confidence vector, which is a combination of objects, attributes, stuff, interactions, relations, and so forth. In this paper [23], the authors presented a semantic attention model based on text conditions. Using this attention model, on the basis of the previously generated text, the encoder can learn on which parts of the image the model needs to focus.

(b) Grid-based visual feature representation without semantic labels: In this type of attention model, the features are extracted using CNN and fed to the attention model. It focuses on those spatial regions that need attention. K. Xu et al. [24], propose an end to end 14×14 VGG-based spatial attention model using hard and soft methods.

(c) Object-based visual representation: In order to infer the latent alignments between image regions and segments of sentences by treating the sentences as weak labels, A. Karpathy et al. [35] presented an alignment model based on Bidirectional RNN (BRNN) and Region-CNN (RCNN). To generate descriptions of the image, they used an end-to-end RNN model.

1.3. Contribution

In our research study, on the basis of preliminary work as reported in [39,40], we exploit the use of VGG16 and InceptionV3 for automatic image captioning. Further, we also introduce a mechanism called attention mechanism to caption non-visual words present in an image and enhance the performance of image captioning. When comparing with the previous work [39,40], the key contribution of this paper is briefly summarized below:

1. In this work, we propose to use two deep learning architectures called VGG16 and InceptionV3 followed by an attribute-based visual representation attention mechanism.
2. Based on this comparative analysis, we investigate how attention mechanisms can be used to find non-visual words present in the image and thus improve the overall image captioning performance.
3. This paper provides an extensive analysis to compare the architecture difference between two pre-trained deep learning models viz VGG16 and InceptionV3 for image captioning.
4. We evaluate the two deep learning architectures on a famous challenging dataset Flicker8k dataset.

The remainder of this section is as: Section 1 gives the introduction along with the motivation, literature survey, and contribution. Section 2 describes the proposed methodology, two pre-trained models, and the Bahdanau attention mechanism, Section 3 discusses the experiment setup, and Sections 4 and 5 provide results and discussion. The conclusion is written in Section 6.

2. Proposed Methodology

To improve caption prediction, we use CNN which produces the best results for image processing. The proposed methodology is shown in Figure 2. CNN is used to extract the features from an image by flattening it. An accurate description of an image requires a comprehensive understanding of visual and non-visual objects. The various visual and non-visual objects have different mutual relationships in different regions of an image. They possess these relationships selectively according to each generated word. It is specifically designed for generating visual as well as non-visual semantic relationships at a fine-grained level. The feature vector is converted into weights for each region of an image, as shown in Figure 2. Each filter kernel of a convolutional layer serves as a semantic detector in CNN. The semantic detector calculates the weighted sum of all the features of

visual and non-visual objects. Semantic knowledge representation provides a semantic relationship among various objects. The semantic information is provided as the input to the attention mechanism. It thus helps in enhancing the performance accuracy of the attention mechanism.

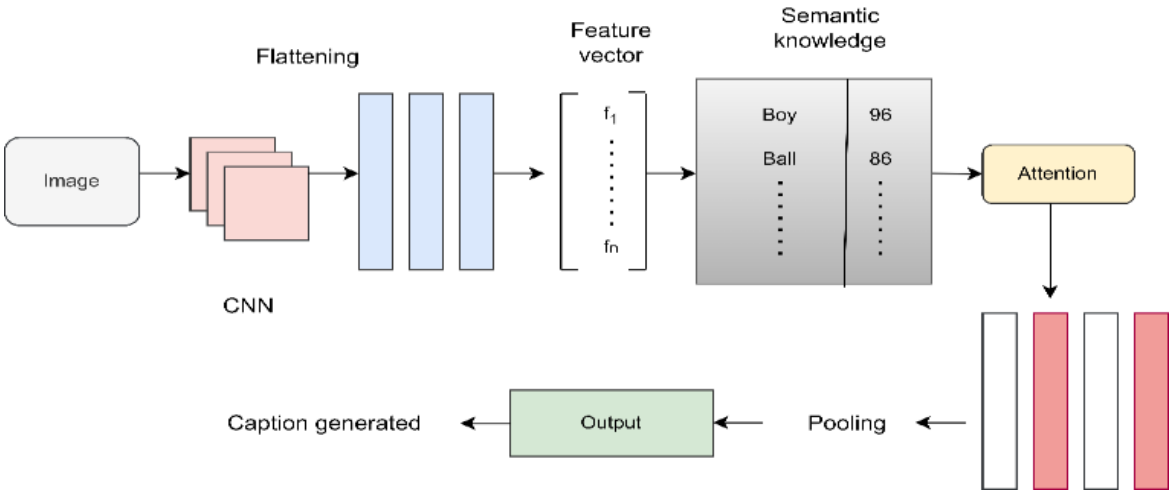


Figure 2. Proposed framework.

Afterwards, high level semantic knowledge is gained. Using the attention mechanism, more focus is given to the important portions. After learning the features, pooling is then applied, and captions are generated. Images have complex features; therefore, CNN employs a convolution layer for extracting features. In order to reduce the computation cost, we used pre-trained CNN models that have already been used on the image datasets. Transfer learning is appropriate for updating models with minimal effort. We can update the models for new data after we have trained them to be highly accurate. Based on our research of literature, we identified two important pre-trained deep learning models. One model is VGG16 [40] and InceptionV3 [39]. For generating captions by deep learning models, the two most important components are a feature extractor and an attention mechanism. Therefore, in our work, we have compared the architecture of two deep-learning pre-trained models and show that it is necessary to have the deep layered architecture for captioning images.

Both the deep learning architectures (VGG16 and InceptionV3) can extract high level features from an image and generate captions; however, the main aim of this study is to examine how the architecture of the model affects the performance of caption generation and if deep layered architecture is necessary for maintaining the performance. By using a pre-trained deep neural network with an attention mechanism, we aim to achieve the features of non-visual objects (like "here", and "with") that play a significant role in giving semantic meaning to the captions of an image. Once a deep neural network is pre-trained using a large database of images for extracting high level features, there is no need to retrain these networks even in the case of image captioning. Therefore, pre-trained models are expected to enhance the computation efficiency and reduce the cost of generating captions. In addition to this, the attention mechanism is used for captioning images as a kind of preference. In the attention mechanism, more focus is given to non-visual words than visual words.

2.1. Deep Feature Extractor

Various classical methods like the histogram of oriented gradients (HOG) and Scale Invariant Feature Transform (SIFT) have been widely employed for many computer vision tasks such as object detection, classification, segmentation, and image retrieval. With the growth of a large amount of collected datasets over the last decade, more and more learning-based descriptors, such as AlexNet, ResNet [26], and GoogleNet [41], have emerged. Unlike

machine learning feature extractors, which require expert knowledge, deep learning models require an abundant amount of training data and computational cost. Despite the fact that various deep learning architectures have been proposed and implemented, the basic principle remains the same: deep learning is a feature representation-learning technology based on a large number of raw image data. It enables a computational model with several neural network layers to learn data representations at various degrees of abstraction. Following the input layer, each layer modifies the preceding layer's representation into a more abstract representation. After learning the discriminative features layer by layer, it will be able to implicitly collect high level features from data on a large scale and use it to represent the original image; however, it will cost a large amount of GPU, time, and processing computation. Therefore, transfer learning using pre-trained models is expected to solve these problems. Deep learning has advanced nearly the whole discipline of computer vision, as well as adjacent areas such as natural language processing and medical image study. Inspired by the success of CNN in the image captioning domain, various pre-trained models are introduced in this domain. However, comparing the architecture of different pre-trained deep learning models with the Bahdanau attention mechanism for image captioning has not been explored yet. Thus, to generate captions for images, two pre-trained models, VGG16 and InceptionV3, have been employed in this work. In this section, the basic architecture and the process of pre-trained models for image captioning are summarized.

2.1.1. Basic VGG16 Architecture for Image Captioning

One of the most commonly used versions of a pre-trained CNN model is VGG16. It consists of 16 layers, of which 13 are convolutional, 2 are fully connected layers, and 1 is a Softmax activation layer [29,42] as shown in Figure 3. To improve the nonlinearity in the model, ReLU (rectified linear unit) activation function is used, and for classification, Softmax is used.

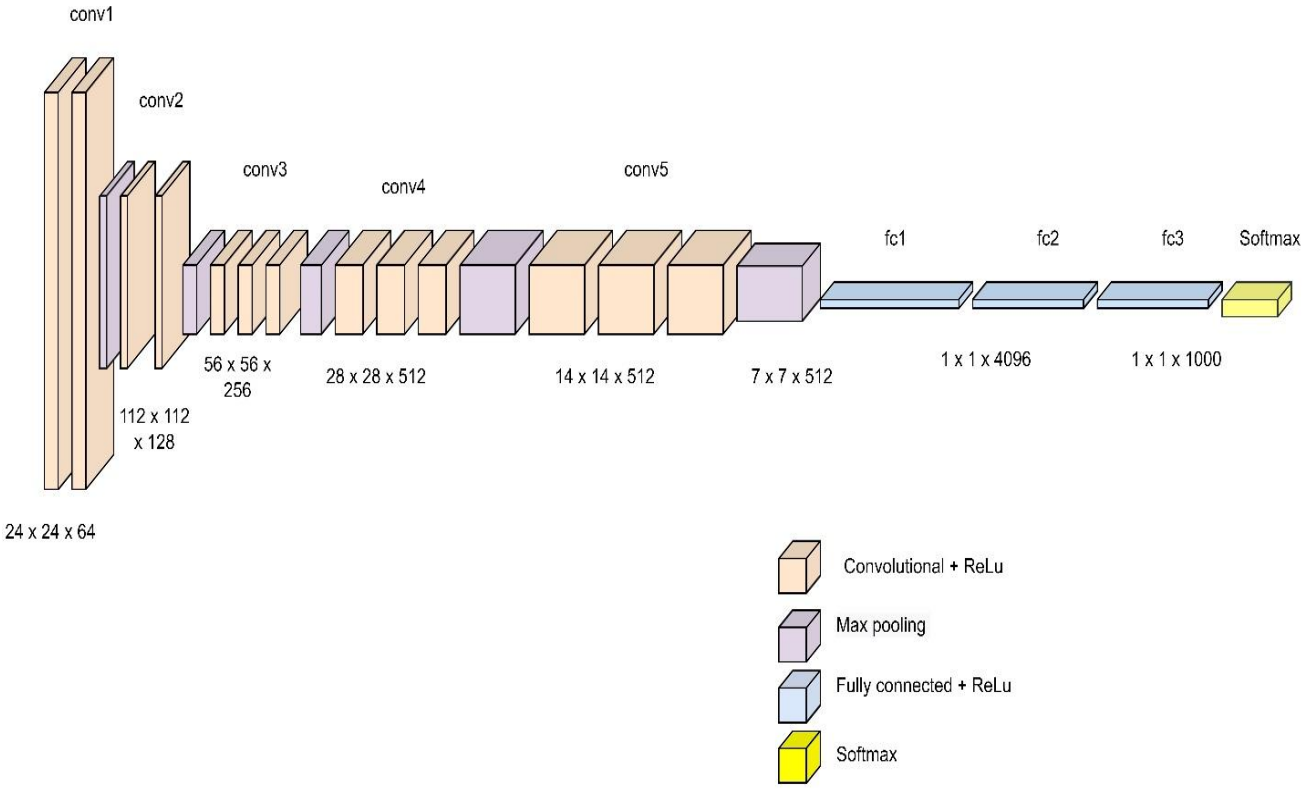


Figure 3. The architecture of VGG16.

The main key aspect of using VGG16 is the small size of the kernel having a stride of 1 pixel and homogenous topology. The total number of parameters is 138 million. VGG16 network configurations consider an input to be a size of 224 × 224 image with three channels—red, green, and blue. The only pre-processing performed is to normalize the RGB values for each pixel, which is accomplished by removing the average value from each pixel. After ReLU activations, the image is sent through the first stack of two convolution layers having sizes 3 × 3. Each layer contains 64 filters. The convolution stride is set to 1 pixel to maintain spatial resolution after convolution. The VGG network's hidden layers use ReLU. The ReLU activation function is used by all the hidden layers of the VGG16 network. However, it is very important to reconfigure the basic model of VGG16, as its main goal was feature extraction for classifying images. Therefore, for generating captions, the last layer is removed.

2.1.2. Basic Inception Architecture for Image Captioning

The basic architecture of InceptionV3 is based on GoogLeNet [41]. One of the key aspects of the Inception structure is their use of Lin's "Network in Network" technique [43], which boosted the representational power of neural networks. This led to the reduction in the dimension to 11 convolutions, therefore reducing the computation cost. The Inception architecture was designed to reduce the computational cost of image classification using deep learning [39]. The Inception module generally has three different convolution sizes and one maximum pooling. As shown in Figure 4, the basic architecture of InceptionV3 constitutes of the following inception modules: an average pooling layer having a filter size of 5 × 5 and stride 3, for dimension reduction it has 1 × 1 layered with ReLU, a fully connected layer having 1024 neuron units with ReLU and a dropout layer. After the convolutional operation, the channel is aggregated and then the fusion operator is performed on the output of the previous layer. Therefore, it helps in reducing overfitting and improving the adaptability of the network. In our approach to InceptionV3, we removed the Softmax layer for classification as the aim of this network is image captioning and not classification. The purpose of each inception module is to capture features at different levels. High level features are captured by 5 × 5 convolutional layers, distributed features are captured using 3 × 3 convolutional layers, and low-level features are extracted by max pooling layers.

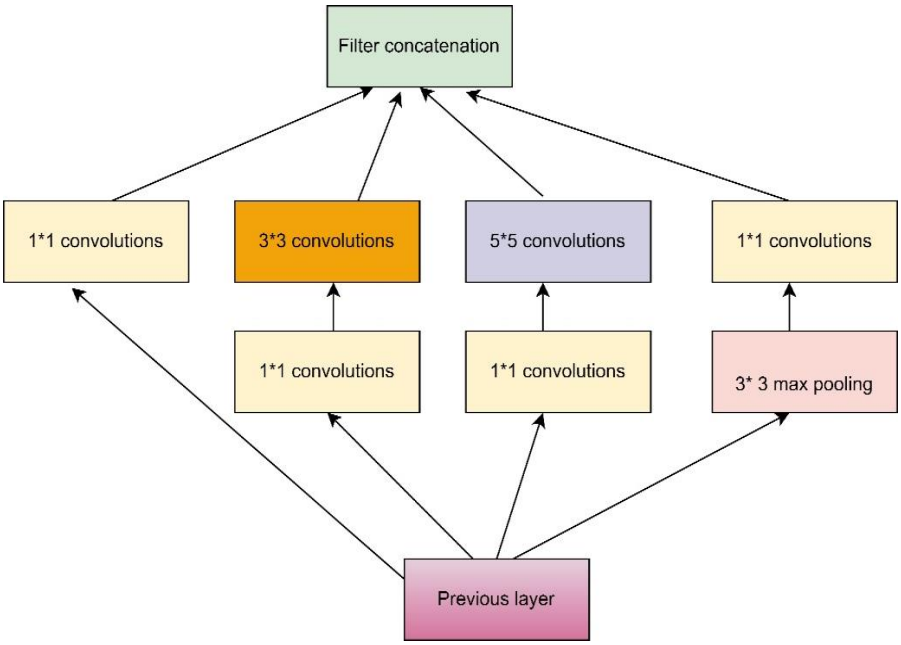


Figure 4. The architecture of InceptionV3.

2.1.3. Transfer Learning

Deep learning is distinguished by the fact that it requires a significant amount of data. Under-fitting will occur during training if the amount of data is insufficient. Researchers introduced the concept of transfer learning to train deep learning networks having limited training data available. With the help of transfer learning, precise and efficient image classification can be achieved [30]. This method is used to pre-train a model on a dataset to learn the features. Practically, the parameters of weight pre-trained on Image Net are then initialized in the model that has been created, guaranteeing that previously learned features are given to your model, resulting in superior outcomes.

2.1.4. Attention Mechanism

Traditional encoder–decoder systems for machine translation encode every source sentence, regardless of length, into a vector of fixed length, from which a translation is generated by a decoder. This made it difficult for the neural network to handle sentence translation, resulting in poor performance. The responsibility of the decoder is to summarize the whole input data into a fixed single vector. However, in some situations, like image captioning, the length of the sentences is too long as well as varying from one image to another. The question to be noted here is whether a fixed vector is capable of capturing all the crucial information related to that image. Moreover, non-visual objects should be given more importance than visual objects.

The goal is to focus on the most relevant words present in the sentence rather than the whole vector. The Bahdanau attention [44] was introduced to address the bottleneck performance of conventional encoder–decoder systems, resulting in significant improvements over the traditional technique. Since it computes linear combinations of encoder and decoder states, therefore it is also called an additive attention mechanism. The basic principle of the Bahdanau attention mechanism is to focus on particular input vectors of the input sequence on the basis of attention weights. Using a set of attention weights, the amount of "attention" to be paid to each input word at each decoding stage is informed to the decoder.

In the case of a non-stacking unidirectional decoder, Bahdanau utilizes the concatenation of the forward and backward hidden states in the bi-directional encoder with the prior target's hidden states. The context vector is generated using all the hidden states of the encoder and the decoder, whereas other models without attention use only the latest encoder hidden state. On the basis of the alignment score represented by a feed-forward neural network, the attention mechanism orients the input and output sequences to pay attention to the most crucial part of the image. On the basis of context vectors related to the source position and the previously generated target words, the model predicts a caption for an image. The architecture of the Bahdanau attention model is shown in Figure 5 and described below:

- The hidden decoder state present at the previous time step $t - 1$ is S_{t-1}.
- A unique context vector c_t at time step t is generated at each decoder step to generate a target word y_t.
- An annotation h_i that captures the crucial information on words focusing around the i-th word out of the total words.
- At the current time step t, the weight value assigned to each annotation h_i is $\alpha_{t,i}$.
- An assigned model a (.) generates an attention score $e_{t,i}$. that shows how well S_{t-1} and h_i matches.

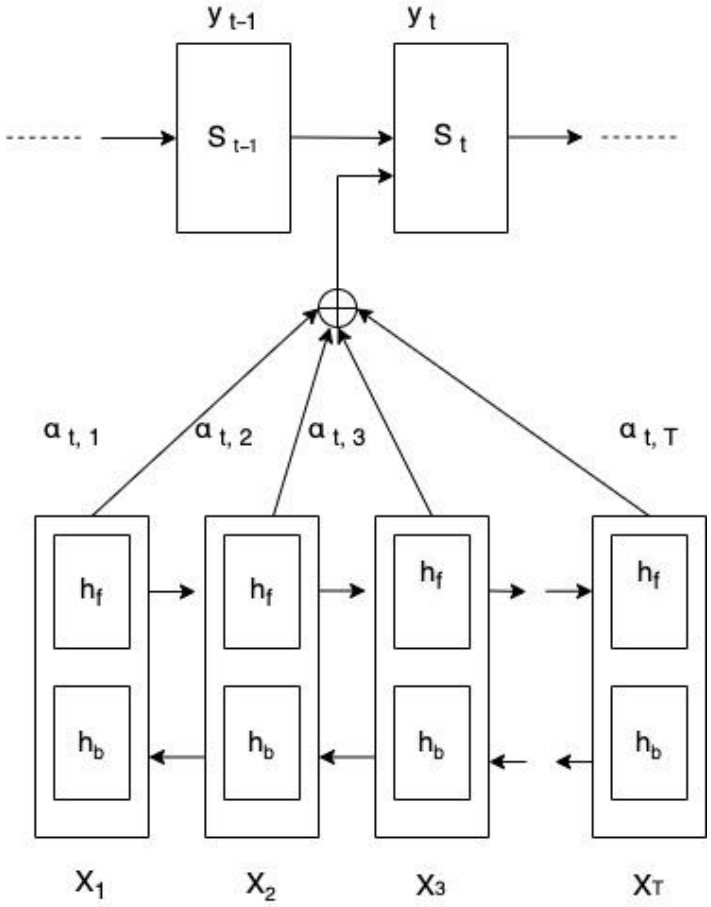


Figure 5. Architecture of Bahdanau with attention mechanism.

The Bahdanau architecture consists of a bidirectional Recurrent neural network (BI-RNN) as encoder and RNN as decoder with an attention mechanism as shown in Figure 5.

The Pseudo code of Bahdanau architecture with attention mechanism is mentioned in Algorithm 1.

Algorithm 1. Pseudo code of Bahdanau architecture with attention mechanism.

1. Using input sentence, the encoder generates a set of annotations, h_i
2. h_i *along with* S_{t-1} (previous hidden decoder state) is fed to a (.); to calculate attention score, i.e., $e_{t,i}$: $e_{t,i} = a(S_{t-1}, h_i)$
3. Apply Softmax function as:

$$\alpha_{t,i} = Softmax(e_{t,i})$$

4. Generate context vector:

$$C_t = \sum_{i=1}^{T} \alpha_{t,i} h_i$$

5. To compute the final output y_t, C_t *and* S_{t-1} *are fed to the decoder.*
6. Repeat steps 2–6 till the end of the sentence.

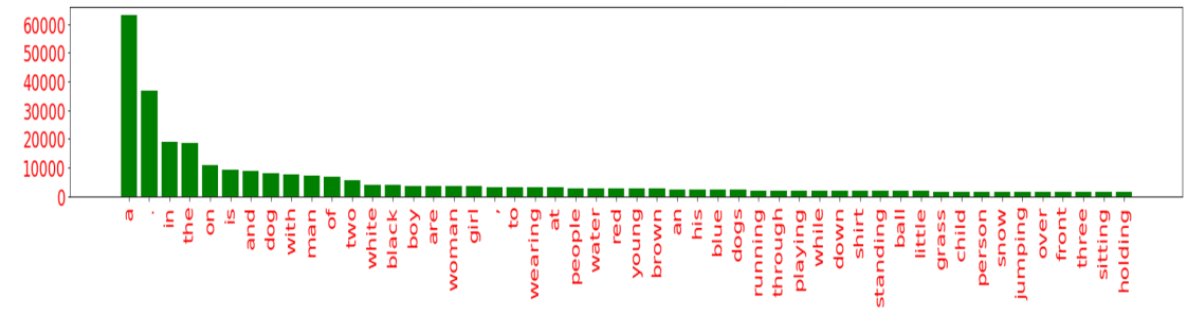
3. Experimental Setup

In the previous section, we discussed the pre-trained deep learning feature extraction techniques. In this section, the dataset and the experiment evaluation are discussed.

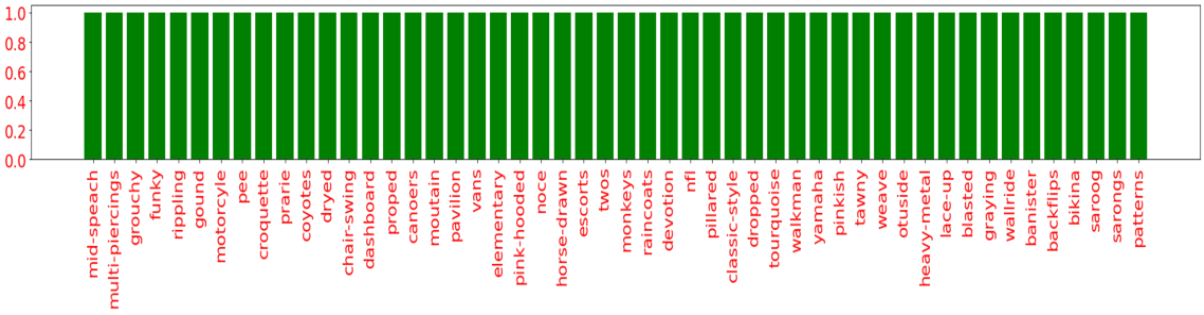
3.1. Dataset

Flicker [32] is the most common dataset for image captioning. It consists of a total of 8092 images in which 6000 are used for training purposes and 2000 for testing the model performance. Each image consists of five related captions. The varied captions are employed for training as well as testing the model. In this work, we filtered the data

by cleaning redundant words from the caption of an image. We first computed the most frequently and less frequently used words. Afterwards, we removed unnecessary words as shown in Figures 6 and 7.

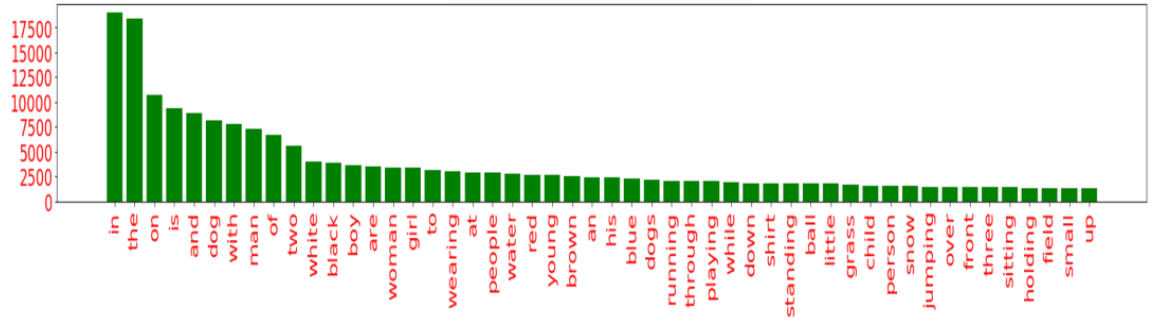


(**a**) The top 50 most frequently appearing words

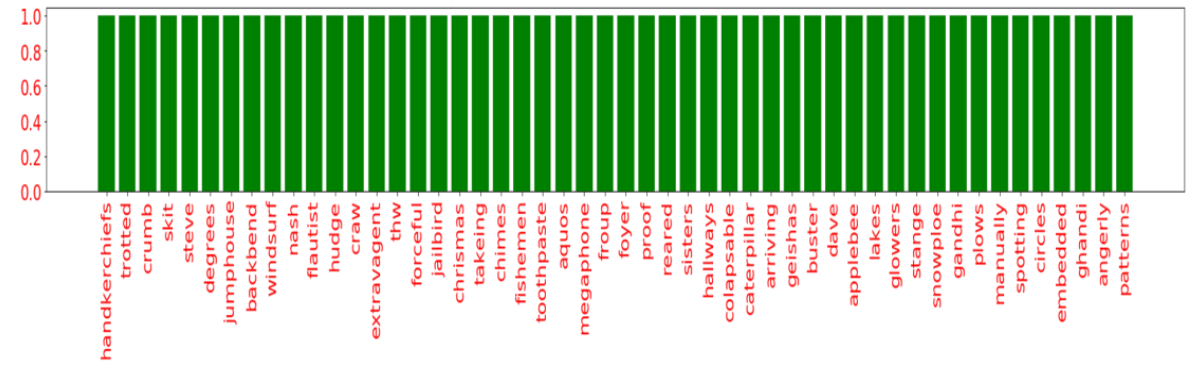


(**b**) The least 50 most frequently appearing words

Figure 6. Top 50 most and least frequently used words in the dataset.



(**a**) The top 50 most frequently appearing words



(**b**) The least 50 most frequently appearing words

Figure 7. Top 50 most and least frequently used words in the dataset after cleaning the captions.

3.2. Performance Metrics

To assess the performance of model prediction, we used various metrics as described below:

1. BLEU (Bilingual evaluation understudy): a well-known machine translation statistic is used to measure the similarity of one sentence with reference to multiple sentences. It was proposed by Papineni et al. in [45]. It returns a value where a higher value represents more similarity. This method works by counting the number of n-grams in one sentence with the n-grams in the reference sentence. A unigram or 1-gram represents a token, whereas a bi-gram indicates each pair of a word. For calculating the BLEU score of multiple sentences, Corpus BLEU is employed, in which a reference list is indicated using a list of documents, and a candidate document is a list where the document is a list of tokens.
2. ROUGE (Recall Oriented Understudy for Gisting Evaluation): It counts the number of "n-grams" that match between the caption generated by our model and a reference. For example, ROUGE-N means using n-grams, etc.
3. METEOR (Metric for Evaluation for Translation with Explicit Ordering): Unlike BLEU, this metric calculates F-score based on mapping unigrams.

3.3. Experimental Setting

We intended to determine the best set of parameters for each model. To achieve that, we divided the data into two sets: training set (6000), testing (1000), and validation set (1000). We first counted the frequency of the most and least frequently used words in our dataset. In the pre-processing step, we added <start> and <end> tags to all the captions. It is done so that the models are able to recognize the start and the end of captions. The next step is to use VGG16, which is a pre-trained model trained on the ImageNet dataset. It consists of convolution as well as fully connected layers. To generate image captions, we performed tokenization and created vocabulary. For each word present in the caption, vector notations are generated.

Encoder and decoder: We implement the VGG16 as encoder and RNN as decoder for generating captions. The last layer of CNN, called Softmax, is used for classification, but in our study, we removed the last layer in order to feed the features to the decoder. For VGG16 to be implemented, the batch size is set to 64, the number of units 512, the learning rate 0.001, and the embedding dimension 256. Adam optimizer is used. The sparse cross-entropy is used to calculate the error. A dropout of 0.5 was used. The total number of parameters used for building the VGG16 model is 138,357,544. Similarly, in order to compare with the InceptionV3 model, we implement the InceptionV3 as encoder and RNN as decoder for generating captions with the same hyperparameters as that of VGG16. However, the total number of trainable parameters of InceptionV3 are 21,768,352.

4. Results

To compare the architecture of two pre-trained deep learning models, we used the Flickr-8k dataset. For this purpose, we divided our experiments into two parts: using VGG16 as an encoder and using InceptionV3 as the encoder. The output of the VGG16 and InceptionV3 encoder is passed on to the RNN decoder. The decoder outputs prediction and the hidden states. To calculate loss, the hidden state is again fed into the model and the predictions are made. BLEU is an algorithm that is used to check the quality of the caption generated. We trained both models for 20 epochs. To reveal the performance of both the ls, we used 6000 images for training and 2000 for testing. The results show that the performance of VGG16 is better than InceptionV3. For image 1, in Table 2 the BLEU score 1 achieved by VGG16 is 62.5, whereas the same score achieved by InceptionV3 is 28.57, as shown in Table 3.

Table 2. Performance metrics obtained using the VGG16 model.

Model	Score	Image 1	Image 2	Image 3
VGG16	BLEU-1	62.5	49.5	10.1
	BLEU-2	42.2	26.7	3.69
	BLEU-3	3	2.1	9.07
	BLEU-4	9.6	7.1	7.06
	ROUGE	0.56	0.60	0.59
	METEOR	0.26	0.31	0.33

Table 3. Performance metrics obtained using InceptionV3 model.

Model	Score	Image 1	Image 2	Image 3
InceptionV3	BLEU-1	28.57	21.9	20.21
	BLEU-2	7.97	5.17	5.22
	BLEU-3	1.75	1.06	1.11
	BLEU-4	1.33	7.95	8.39
	ROUGE	0.59	0.54	0.55
	METEOR	0.27	0.29	0.23

Similarly, in the case of score-2, VGG16 achieved 42.2, whereas InceptionV3 achieved 7.97, which is very small compared to the former one. However, in the case of image 3, InceptionV3 achieved better results than VGG16. One of the interesting facts that we found was that using an attention mechanism helped in focusing on non-visual objects. Using ROUGE metrics, Table 2 demonstrates that VGG16 performs substantially better than InceptionV3. From tabure 8, it is evident that the VGG16 model performed better than the deep InceptionV3 model.

Furthermore, we have compared our results with ResNet using different transfer functions. When comparing the results of VGG16 from Table 2 with the results of ResNet from Table 4, we can see that VGG16 performs better than ResNet. It is because, with an increase in the layers of architecture, the feature extraction capability also increases.

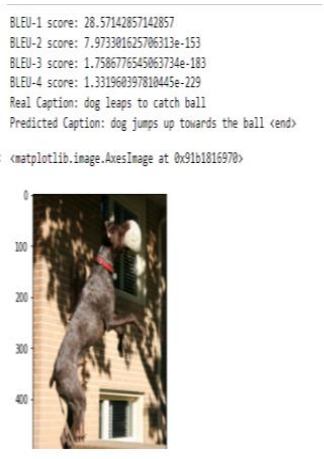
Table 4. Performance metrics obtained using different ResNet models.

S. No.	Model	Score	Image 1
1.	ResNet [23]	BLEU-1	29.5
2.		BLEU-2	30.5
3.		BLEU-3	29
4.		BLEU-4	30
5.		ROUGE	-
6.		METEOR	0.24
7.	ResNet with ReLU [23]	BLEU-1	29
		METEOR	25
		ROUGE	-
8.	ResNet with tanh [23]	BLEU-1	25
		METEOR	25
		ROUGE	-
9.	ResNet with Softmax [23]	BLEU-1	29
		METEOR	24
		ROUGE	-
10.	ResNet with Sigmoid [23]	BLEU-1	30
		METEOR	24
		ROUGE	-

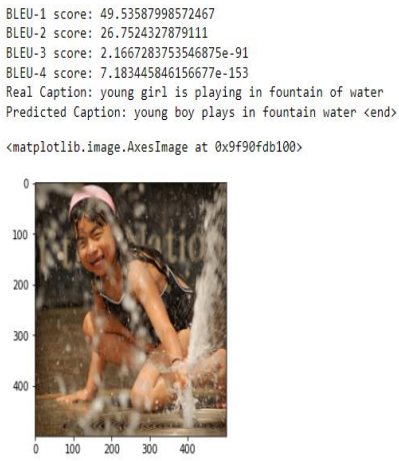
Summarizing all the above results, it is concluded that VGG16 along with the Bahdanau attention mechanism shows better performance than the InceptionV3 model.

Figure 8 shows the results of the VGG16 and InceptionV3 Models. It can be inferred from Figure that VGG16 is outperforming the InceptionV3 model in terms of predicting the caption of the images.

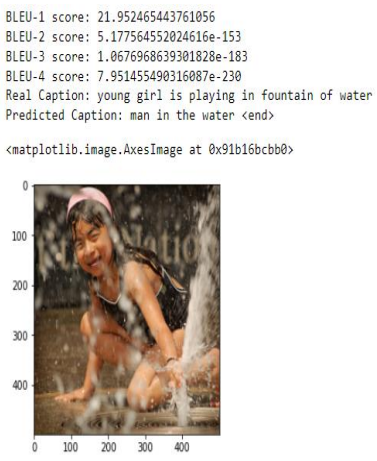
(**a**) Caption predicted by VGG16 for image 1



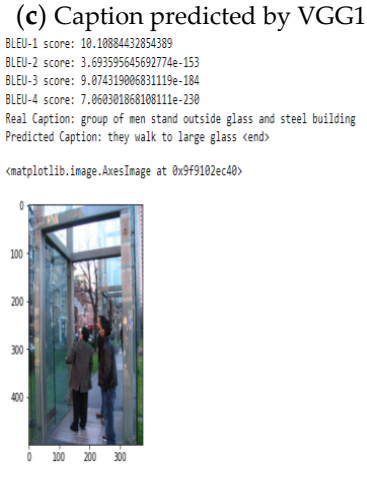
(**b**) Caption predicted by InceptionV3 for image 1



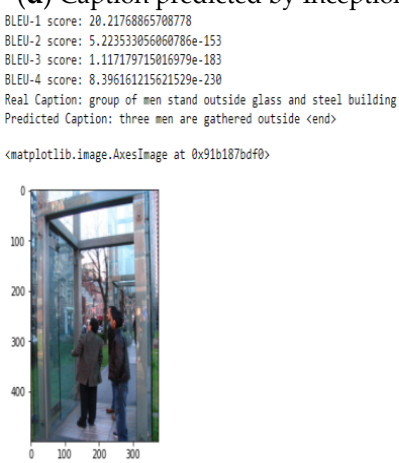
(**c**) Caption predicted by VGG16 for image 2



(**d**) Caption predicted by InceptionV3 for image 2



(**e**) Caption predicted by VGG16 for image 3



(**f**) Caption predicted by InceptionV3 for image 3

Figure 8. Shows visual representation of captions using two different models.

5. Discussion

In the presented work, two different architectures of pre-trained models along with the Bahdanau attention mechanism are utilized. Through our work, we are able to show that deep learning techniques with transfer learning and fine-tuning had a substantial effect on automatic image caption generation. The task of generating a caption for an image can be seen trending from the last few years. Maru et al. in [46] used VGG16 and InceptionV3 as an encoder for generating captions. They concluded that the performance of VGG16 is better than InceptionV3. However, accurate image caption generation requires an understanding of semantic knowledge as well. Therefore, the two main goals of the attention mechanism are to enable a feature extractor model to learn semantic features by aligning together objects in images and words. The second goal is to learn a common feature space, where an image and language can be modeled together. After training our models, we ran them three times on random images to check their performance. Our results support the robustness of transfer learning with an attention mechanism.

When the predicted sentence is compared with one reference, it is called BLEU score 1. Similarly, when it is compared with two sentences, it is called BLEU score 2. Table 2 demonstrates the BLEU score 1 for image 1, 2, and 3 achieved by the VGG16 encoder using Bahdanau attention is far better than that of InceptionV3. This is because of the deeper architecture of VGG16 as compared to that of InceptionV3. The same performance can be seen in the case of other images as well using other metrics like ROUGE and METEOR. The importance of the attention mechanism can be seen in Figure 8. Using the Bahdanau attention mechanism, the model is capable of high diversity. For image 1, the word "up" and "the" has been used by the VGG16 model to predict the caption of an image. In Figure 8a, we can see that VGG 16 predicted the caption for an image accurately by mentioning the term "catch", which cannot be found in Figure 8b. Similarly, for Figure 8c, VGG16 is able to predict more accurate words like "plays" and "fountain", which InceptionV3 fails to (Figure 8d). For Figure 8e, again the prediction of VGG16 is better than InceptionV3 (Figure 8f). It used words like "to", "large" and "glass" to describe the image. In contrast, InceptionV3 described the word very briefly. Therefore, we can say that the caption generated by VGG16 using the Bahdanau attention mechanism is more accurate and detailed. Thus, an efficient automatic image captioning model can be achieved using a pre-trained model like VGG16 along with the attention model. It also proves that transfer learning can be used to generate captions.

6. Conclusions

The objective of this work was to develop an automatic image caption generation model using various CNN. The work done here has an overarching theme of leveraging transfer learning techniques. Therefore, in this paper, we analyzed and compared the performance of two pre-trained deep learning models, i.e., VGG16 and InceptionV3. Further, we incorporated the Bahdanau attention mechanism with these two pre-trained deep learning CNN architectures. Using the Bahdanau attention mechanism, we show how semantic information plays a vital role in generating accurate captions. Experiments were conducted to determine the efficacy of attention-based learning in comparison to traditional image captioning mechanisms. Our research has yielded positive results. Through our results, we depict that the VGG16 encoder with the Bahdanau attention mechanism performs better than the InceptionV3 encoder on the Flicker8k dataset. However, Filckr8K is a small dataset due to which deeper models outperform shallow ones on the retrieval task. In the future, large datasets can be used to see whether shallow models can replace deeper models. Moreover, this work can be expanded by including various attention mechanisms with different pre-trained CNN architectures. New techniques in deep learning like Transformers can be introduced to assess the local and restricted attention mechanism so that the model can generate captions of different types of inputs such as images, videos, and audios, etc.

Author Contributions: Conceptualization, S.A., Y.G., F.A.R. and S.T.; methodology, S.A.; software, S.A. and F.A.R.; validation, Y.G. and F.A.R.; formal analysis, S.A. and F.A.R.; investigation, S.A.; resources, Y.G. and S.T.; writing—original draft, S.A.; writing—review and editing, Y.G., F.A.R. and S.T.; visualization, S.A.; supervision, Y.G. and S.T.; project administration, Y.G. and S.T.; funding acquisition, Y.G. and S.T. All authors have read and agreed to the published version of the manuscript.

Funding: The authors would like to thank the United Arab Emirates University for funding this work under UAEU Strategic Research Grant G00003676 (Fund No.: 12R136) through Big Data Analytics Center.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The dataset used in this study is a public dataset Flicker [32].

Acknowledgments: This work was supported by the Deanship of Scientific Research, Vice Presidency for Graduate Studies and Scientific Research, King Faisal University, Saudi Arabia, under Project GRANT2,086 and United Arab Emirates University for funding this work under UAEU Strategic Research Grant G00003676 (Fund No.: 12R136) through Big Data Analytics Center.

Conflicts of Interest: The authors declare that there is no conflict of interest regarding the publication of this paper.

References

1. Wang, P.; Yang, A.; Men, R.; Lin, J.; Bai, S.; Li, Z.; Ma, J.; Zhou, C.; Zhou, J.; Yang, H. OFA: Unifying Architectures, Tasks, and Modalities Through a Simple Sequence-to-Sequence Learning Framework. *arXiv* **2022**, arXiv:2202.03052. Available online: https://arxiv.org/abs/2202.03052 (accessed on 14 July 2022).
2. Hsu, T.Y.; Giles, C.L.; Huang, T.H. SCICAP: Generating Captions for Scientific Figures. In *Findings of the Association for Computational Linguistics, Findings of ACL: EMNLP 2021*; Association for Computational Linguistics: Punta Cana, Dominican Republic, 2021; pp. 3258–3264. [CrossRef]
3. Hossain, M.Z.; Sohel, F.; Shiratuddin, M.F.; Laga, H.; Bennamoun, M. Text to Image Synthesis for Improved Image Captioning. *IEEE Access* **2021**, *9*, 64918–64928. [CrossRef]
4. Sehgal, S.; Sharma, J.; Chaudhary, N. Generating Image Captions Based on Deep Learning and Natural Language Processing. In Proceedings of the ICRITO 2020 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) IEEE, Noida, India, 4–5 June 2020; pp. 165–169. [CrossRef]
5. Jain, H.; Zepeda, J.; Perez, P.; Gribonval, R. Learning a Complete Image Indexing Pipeline. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 4933–4941. [CrossRef]
6. Pang, S.; Orgun, M.A.; Yu, Z. A Novel Biomedical Image Indexing and Retrieval System via Deep Preference Learning. *Comput. Methods Prog. Biomed.* **2018**, *158*, 53–69. [CrossRef] [PubMed]
7. Makav, B.; Kilic, V. A New Image Captioning Approach for Visually Impaired People. In Proceedings of the 11th International Conference on Electrical and Electronics Engineering (ELECO 2019), Bursa, Turkey, 28–30 November 2019; pp. 945–949. [CrossRef]
8. Zhang, Z.; Wu, Q.; Wang, Y.; Chen, F. High-Quality Image Captioning with Fine-Grained and Semantic-Guided Visual Attention. *IEEE Trans. Multimed.* **2019**, *21*, 1681–1693. [CrossRef]
9. Alam, S.; Raja, P.; Gulzar, Y. Investigation of Machine Learning Methods for Early Prediction of Neurodevelopmental Disorders in Children. *Wirel. Commun. Mob. Comput.* **2022**, *2022*, 5766386. [CrossRef]
10. Sahlan, F.; Hamidi, F.; Misrat, M.Z.; Adli, M.H.; Wani, S.; Gulzar, Y. Prediction of Mental Health Among University Students. *Int. J. Perceptive Cogn. Comput.* **2021**, *7*, 85–91.
11. Khan, S.A.; Gulzar, Y.; Turaev, S.; Peng, Y.S. A Modified HSIFT Descriptor for Medical Image Classification of Anatomy Objects. *Symmetry* **2021**, *13*, 1987. [CrossRef]
12. Gulzar, Y.; Khan, S.A. Skin Lesion Segmentation Based on Vision Transformers and Convolutional Neural Networks—A Comparative Study. *Appl. Sci.* **2022**, *12*, 5990. [CrossRef]
13. Albarrak, K.; Gulzar, Y.; Hamid, Y.; Mehmood, A.; Soomro, A.B. A Deep Learning-Based Model for Date Fruit Classification. *Sustainability* **2022**, *14*, 6339. [CrossRef]
14. Gulzar, Y.; Hamid, Y.; Soomro, A.B.; Alwan, A.A.; Journaux, L. A Convolution Neural Network-Based Seed Classification System. *Symmetry* **2020**, *12*, 2018. [CrossRef]
15. Hamid, Y.; Wani, S.; Soomro, A.B.; Alwan, A.A.; Gulzar, Y. Smart Seed Classification System Based on MobileNetV2 Architecture. In Proceedings of the 2nd International Conference on Computing and Information Technology, ICCIT 2022, Tabuk, Saudi Arabia, 25–27 January 2022; pp. 217–222. [CrossRef]
16. Hamid, Y.; Elyassami, S.; Gulzar, Y.; Balasaraswathi, V.R.; Habuza, T.; Wani, S. An Improvised CNN Model for Fake Image Detection. *Int. J. Inf. Technol.* **2022**, 1–11. [CrossRef]

17. Faris, M.; Hanafi, F.M.; Sukri Faiz, M.; Nasir, M.; Wani, S.; Abdulkhaleq, R.; Abdulghafor, A.; Gulzar, Y.; Hamid, Y. A Real Time Deep Learning Based Driver Monitoring System. *Int. J. Perceptive Cogn. Comput.* **2021**, *7*, 79–84.
18. Sharma, H.; Jalal, A.S. Incorporating External Knowledge for Image Captioning Using CNN and LSTM. *Mod. Phys. Lett. B* **2020**, *34*, 2050315. [CrossRef]
19. Wang, C.; Yang, H.; Bartz, C.; Meinel, C. Image Captioning with Deep Bidirectional LSTMs. In Proceedings of the 2016 ACM Multimedia Conference, Amsterdam, The Netherlands, 15–19 October 2016; pp. 988–997. [CrossRef]
20. Aneja, J.; Deshpande, A.; Schwing, A.G. Convolutional Image Captioning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 5561–5570. [CrossRef]
21. Yang, X.; Zhang, H.; Cai, J. Learning to Collocate Neural Modules for Image Captioning. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 10–17 October 2019; pp. 4249–4259. [CrossRef]
22. Khan, R.; Islam, M.S.; Kanwal, K.; Iqbal, M.; Hossain, M.I.; Ye, Z. A Deep Neural Framework for Image Caption Generation Using GRU-Based Attention Mechanism. *arXiv* **2022**, arXiv:2203.01594.
23. Zhou, L.; Xu, C.; Koch, P.; Corso, J.J. Watch What You Just Said: Image Captioning with Text-Conditional Attention. In Proceedings of the Thematic Workshops 2017—Proceedings of the Thematic Workshops of ACM Multimedia 2017, Co-Located with MM 2017, Mountain View, CA, USA, 23–27 October 2017; pp. 305–313. [CrossRef]
24. Xu, K.; Ba, J.L.; Kiros, R.; Cho, K.; Courville, A.; Salakhutdinov, R.; Zemel, R.S.; Bengio, Y. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. In Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6–11 July 2015; Volume 3, pp. 2048–2057.
25. Lu, J.; Xiong, C.; Parikh, D.; Socher, R. Knowing When to Look: Adaptive Attention via a Visual Sentinel for Image Captioning. In Proceedings of the 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, 21–26 July 2017; pp. 3242–3250. [CrossRef]
26. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778. [CrossRef]
27. Anderson, P.; He, X.; Buehler, C.; Teney, D.; Johnson, M.; Gould, S.; Zhang, L. Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 6077–6086. [CrossRef]
28. Yao, T.; Pan, Y.; Li, Y.; Mei, T. Exploring Visual Relationship for Image Captioning. In *Computer Vision—ECCV 2018, 15th European Conference, Munich, Germany, 8–14 September 2018*; Part XIV; Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics); Springer: Berlin/Heidelberg, Germany, 2018; Volume 11218, pp. 711–727. [CrossRef]
29. Chen, F.C.; Jahanshahi, M.R. NB-CNN: Deep Learning-Based Crack Detection Using Convolutional Neural Network and Naïve Bayes Data Fusion. *IEEE Trans. Ind. Electron.* **2018**, *65*, 4392–4400. [CrossRef]
30. Gupta, R.; Bhardwaj, K.K.; Sharma, D.K. Transfer Learning. In *Machine Learning and Big Data: Concepts, Algorithms, Tools and Applications*; John Wiley & Sons: Hoboken, NJ, USA, 2020; pp. 337–360. [CrossRef]
31. Huang, L.; Wang, W.; Chen, J.; Wei, X.Y. Attention on Attention for Image Captioning. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 4633–4642. [CrossRef]
32. Hodosh, M.; Young, P.; Hockenmaier, J. Framing Image Description as a Ranking Task: Data, Models and Evaluation Metrics. In Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence (IJCAI 2015), Buenos Aires, Argentina, 25–31 July 2015; pp. 4188–4192.
33. Vinyals, O.; Toshev, A.; Bengio, S.; Erhan, D. Show and Tell: A Neural Image Caption Generator. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 3156–3164. [CrossRef]
34. Chen, L.C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *40*, 834–848. [CrossRef]
35. Karpathy, A.; Li, F.-F. Deep Visual-Semantic Alignments for Generating Image Descriptions. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 664–676. [CrossRef]
36. Li, L.; Tang, S.; Zhang, Y.; Deng, L.; Tian, Q. GLA: Global-Local Attention for Image Description. *IEEE Trans. Multimed.* **2018**, *20*, 726–737. [CrossRef]
37. Ding, G.; Chen, M.; Zhao, S.; Chen, H.; Han, J.; Liu, Q. Neural Image Caption Generation with Weighted Training and Reference. *Cogn. Comput.* **2019**, *11*, 763–777. [CrossRef]
38. Yan, S.; Xie, Y.; Wu, F.; Smith, J.S.; Lu, W.; Zhang, B. Image Captioning via Hierarchical Attention Mechanism and Policy Gradient Optimization. *Signal Process.* **2020**, *167*, 107329. [CrossRef]
39. Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the Inception Architecture for Computer Vision. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 2818–2826. [CrossRef]
40. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. In Proceedings of the 3rd International Conference on Learning Representations, ICLR 2015—Conference Track Proceedings, San Diego, CA, USA, 7–9 May 2015; pp. 1–14. [CrossRef]

41. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going Deeper with Convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 1–9. [CrossRef]
42. Seo, J.; Han, S.; Lee, S.; Kim, H. Computer Vision Techniques for Construction Safety and Health Monitoring. *Adv. Eng. Inform.* **2015**, *29*, 239–251. [CrossRef]
43. Lin, M.; Chen, Q.; Yan, S. Network in Network. In Proceedings of the 2nd International Conference on Learning Representations, ICLR 2014—Conference Track Proceedings, Banff, AB, Canada, 14–16 April 2014; pp. 1–10.
44. Bahdanau, D.; Cho, K.H.; Bengio, Y. Neural Machine Translation by Jointly Learning to Align and Translate. In Proceedings of the 3rd International Conference on Learning Representations, ICLR 2015 Conference Track Proceedings, San Diego, CA, USA, 7–9 May 2015; pp. 1–15.
45. Papineni, K.; Roukos, S.; Ward, T.; Zhu, W.J. Bleu: A method for automatic evaluation of machine translation. In Proceedings of the 40th annual meeting of the Association for Computational Linguistics, Philadelphia, PA, USA, 7–12 July 2002; pp. 311–318.
46. Maru, H.; Chandana, T.S.S.; Naik, D. Comparison of Image Encoder Architectures for Image Captioning. In Proceedings of the 5th International Conference on Computing Methodologies and Communication, ICCMC 2021, Erode, India, 8–10 April 2021; pp. 740–744. [CrossRef]