Automatic Image Caption Generation Using Deep Learning

Akash Verma
National Institute of Technology Hamirpur

Arun Kumar Yadav
National Institute of Technology Hamirpur

Mohit Kumar
National Institute of Technology Hamirpur

Divakar Yadav (✉ dsy99@rediffmail.com)
National Institute of Technology https://orcid.org/0000-0001-6051-479X

Research Article

Keywords: Image, Neural Network, Caption, CNN, Feature Extraction, RNN, LSTM

Posted Date: June 21st, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1282936/v1

Automatic Image Caption Generation Using Deep Learning

Akash Verma[1], Arun Kumar Yadav[1], Mohit Kumar[1] and Divakar Yadav[1*]

[1*]Department of CSE, NIT Hamirpur, Hamirpur, 177005, Himachal Pradesh, India.

*Corresponding author(s). E-mail(s): dsy99@rediffmail.com;

Abstract

Image captioning is an interesting and challenging task with applications in diverse domains such as image retrieval, organizing and locating images of users' interest etc. It has huge potential for replacing manual caption generation for images and is especially suitable for large scale image data. Recently, deep neural network based methods have achieved great success in the field of computer vision, machine translation and language generation. In this paper, we propose an encoder-decoder based model that is capable of generating grammatically correct captions for images. This model makes use of VGG16 Hybrid Places 1365 as encoder and LSTM as decoder. To ensure the complete ground truth accuracy, the model is trained on the labelled Flickr8k and MSCOCO Captions datasets. Further, the model is evaluated using all standard metrics such as BLEU, METEOR, GLEU and ROUGE_L. Experimental results indicate that the proposed model obtained a BLEU-1 score 0.6666, METEOR score 0.5060 and GLEU score 0.2469 on Flickr8k dataset and BLEU-1 score 0.7350, METEOR score 0.4768 and GLEU score 0.2798 on MSCOCO Captions dataset. Thus, the proposed method achieved a significant performance as compared to the state-of-art approaches. To evaluate the efficacy of the model further, we also show the results of caption generation from live sample images that reinforces the validity of the proposed approach.

Keywords: Image, Neural Network, Caption, CNN, Feature Extraction, RNN, LSTM

1 Introduction

Automatic image caption generation is an active research topic in computer vision. Researchers are drawn to this topic because it has a wide range of practical applications and combines two main AI fields, Natural language processing & Computer vision. To make a meaningful phrase out of an image, we first need to understand it, that we can do with the help of image classification and object identification [52].

In fact, the task of automatic generation of image caption is harder than object detection and image classification. Human beings are able to generate the caption of images by viewing them. The ability to locate visual objects that are depicted in an image, along with ascertaining the association betwixt those images, comes naturally to human beings. The ability to generate caption is achieved with experience and learning. It is hypothesized that machines can achieve this ability by training on different type of datasets to understand the relation between objects and achieve accuracy similar to human beings.

In light of enormous developments in the field of Artificial Intelligence (AI), images are being used as input for various tasks.One of the application of AI has been discussed in the paper [27]. They used deep learning methods to recognize the face. The main objective of automatic image caption generation is to generate well formed sentence that describe content of the image and relation between the objects detected from the image, that may be used for recommendations in various applications. It may be used for visually impaired persons, in virtual assistants, image indexing, social media recommendation, and for several other natural language processing applications [10, 5]. Image caption generation can help machine to understand the image content. It is not only the process of detecting the objects in an image but also understanding the relation between the detected objects. Researchers categorize image captioning methods into Template-based, Retrieval-based and Deep neural network based methods [36]. In template based methods, first attributes, objects and actions are detected from the image and then predefined templates with number of blank slots are filled. In retrieval based methods, caption is generated by retrieving an image that is similar to the input image. These methods generate syntactically correct captions, although image specificity and semantic correctness is not guaranteed. In Deep neural network based methods, first image is encoded and then captions are generated using language model. Compared to the first two methods, deep neural networks based method may be able to generate semantically more accurate captions for the given images. Deep neural network architecture is used in the majority of existing image captioning models. Kiros et al. [28] were the first to use this type of model to define a multi-modal log-bilinear model for image captioning with a fixed context window. In encoder-decoder image captioning models mostly a CNN and a RNN are used. CNN is used as encoder to convert the input image into 1-D array representation, and RNN as decoder or language model to generate the caption. Identification of proper CNN and RNN models is a challenging issue. The summary of our research contributions is as follows:

- This study proposes VGG16 Hybrid Places 1365 and LSTM as encoder and decoder respectively for automatic image captioning. The proposed model outperforms all state-of-art approaches.
- This study reports experimental results using all popular metrics such as BLEU, ROUGE_L, METEOR and GLEU on Flickr8k and MSCOCO Captions datasets.
- The study reports results of the proposed model on random live images for validation.

In regard to contribution 1, the present study has employed VGG16 Hybrid Places 1365 model as the encoder. This model is trained on both the ImageNet (1000 classes) and the Places 365 (365 classes) datasets. Thus, the total number of output classes tackled are 1365. It is expected that the larger number of samples and classes in the training set will lead to superior generalizability and provide more accurate results on the image caption generation task. This is in consonance with the obtained results from the experiments. In regard to contribution 2, there are several researches in literature that have used only some specific metrics out of all the available choices [34, 49, 33, 24, 3, 15, 10, 25]. This could potentially lead to unfair evaluation of the results. Therefore, in order to avoid such a situation, the current study reports the results in all popular metrics. In order to demonstrate the efficacy of the proposed model, it is evaluated for 10 sample images from the dataset along with their reference captions (ground-truth) in Figure 6. The result on these random images further strengthens the claimed effectiveness of the model.

The motivations regarding the novelty of the proposed model are as follows. Firstly, before settling on the proposed model, several other models and their hybrid variants were evaluated in initial experiments on publicly available Flickr8k dataset. Based on those initial results, it was decided to work upon the method that is proposed in this article. The proposed model provides state of art results for image captioning on public datasets such as Flickr8k and MSCOCO Captions. The superior results on two datasets show the generalizability of the proposed model. Secondly, the generated captions through the proposed on live sample images again reaffirms the efficacy of the proposed model. Thirdly, in most of the studies in literature, the authors have used only specific evaluation metrics to report results. But such specific metrics do not present a fair evaluation of their respective approaches. In contrast, the proposed model is evaluated on all possible metrics to show a fair and just evaluation. As can be seen from the results, the proposed model provides a balanced performance on all metrics, thereby demonstrating the efficacy and novelty of the model.

The rest of the paper is laid out in the following manner. Section 2 describes the related research in image captioning. Section 3 discusses the background concepts and the proposed model along with its evaluation parameters. Section 4 describes the results of proposed model and analyses the results, followed by conclusions and future scope in Section 5.

2 Related Literature Review

A wide range of research has been carried out previously in the field of image captioning and content generation for image captioning.Tn the paper [4], authors proposed content selection methods for image caption generation. In this paper, Mainly three categories of features i.e geometric, conceptual, and visual are used for content generation.Also, variety of methods have been proposed for image caption generation in the past. They may be classified in broadly three categories i.e., Template-based methods [14, 30, 51, 37], Retrieval-based methods [40, 43, 16, 46, 20], and Deep neural network based (Encoder-decoder) methods [49, 3, 15, 12]. These models are often built using CNN to encode the image & extract visual information whereas RNN is used to decode the visual information into a sentence. A detail study has been done on deep neural network-based image captioning models in the paper [32]. In this paper, authors surveyed the deep learning based models used for image captioning on MS-COCO and Flickr30k dataset.

In template based approach (Fig.1 (a)), the process of caption generation is performed using predefined templates with a number of blank spaces that are filled with objects, actions, and attributes recognised in the input image. In the paper [14], the authors proposed the template slots for generating captions that are filled with the predicted triplet (object, action, scene) of visual components. Again, in the paper [30], the authors used a Conditional Random Field (CRF) based technique to derive the objects (people, cars etc. or things like trees, roads etc.), attributes, and prepositions. The model is evaluated on PASCAL dataset using BLEU and ROUGE scores. In this work, the best BLEU score obtained was 0.18 and corresponding best ROUGE score was 0.25. The authors of the paper [9] presented a method for caption generation by selecting valuable phrases from existing captions and combines them carefully to create a new caption. It used 1 million captioned pictures corpus dataset, with 1000 images put aside as a test set to compute BLEU (0.189) and METEOR (0.101) score. These methods are basically hard to design and depend on pre-defined template. Due to their dependence on such tempates, these methods are not able to generate sentences or captions with variable lengths. Template-based approaches are capable of generating captions that are grammatically correct. However, the length of the generated captions is fixed due to the predefined templates. In the paper [8], authors proposed a memory-enhanced captioning model for image captioning. They introduce external memory based past knowledge to encode the caption, and further use the decoder to generate correct caption. They evaluate the propose model on MS-COCO dataset with 3.5% improved CIDr.

In retrieval based captioning methods (Fig.1 (b)), the captions for images are generated by collecting visually similar images. This type of approaches find captions for visually comparable images from the training dataset after discovering visually similar images and use those captions to return the caption of the query image. On the basis of millions of photos and their descriptions, the authors of the paper [40] developed a model for finding similar images among the large number of images in the dataset and returning the descriptions of these retrieved images to query image. In the paper [35], the authors used density estimation method to generate captions and got a BLEU score of approximately 0.35. Again, in the paper [46], the authors used visual and semantic similarity scores to cluster similar images. They merge the images together, retrieve caption of the input image from captions of similar images in the same cluster. Some researchers have proposed a ranking-based framework to generate captions for each image by exposing it to sentence-based image captioning [20]. In the paper [18], authors proposed text based visual attention (TBVA) model for identifying salient object automatically. They evaluated proposed model on MS-COCO and Flickr30k dataset.In the paper [39], authors proposed data-driven based approach for image description generation using retrieval based technique. They concluded that proposed method provide efficient and relevant result to produce image captions. Although these strategies provide syntactically valid and generic sentences, they fail to produce image-specific and semantically correct sentences.

Due to the success of encoder-decoder architectures (Fig.1 (c)) in machine translation, a similar encoder-decoder (neural network based method) architecture has been successfully used in image captioning as well. These models depend on deep neural networks for producing description of input images, that are considered more precise than those generated by the other two categories of methods. In the paper[13],authors proposed dual graph convolutional networks with transformer and curriculum learning for image captioning. They evaluated results on MS-COCO dataset with achieves BLEU-1 score of 82.2 and a BLEU-2 score of 67.6. The authors of the paper [49] proposed the NIC (Neural Image Caption) model based on encoder-decoder architecture. In this model, CNN is utilised as the encoder, and the final layer of CNN is linked to the RNN decoder, which generates the text captions. In this model, LSTM is utilized as RNN. Junhua Mao et al. [33] proposed the m-RNN model to accomplish the task of generating captions and the task of image and sentence retrieval. This model provided a BLEU-1 score of 0.5650 on the Flickr8k dataset. Chetan Amritkar et al. [3] designed a neural network model which generates captions for images in natural language and yielded a score of 0.5335 (BLEU-1) on Flickr8k dataset. In the method of [50], an image is encoded into a numerical representation using convolutional neural network (CNN) and the output of CNN is used as input in the decoder (RNN) to generate the captions but one word at a time. Yan Chu et al. [10] put forward a model using ResNet50 and LSTM with soft attention that produced a BLEU-1 score of 0.619 on the Flickr8K dataset. Sulabh Katiyar et al. [45] proposed two types of models,a simple encoder-decoder model and an encoder-decoder model with attention. These models generate BLEU-1 scores of 0.6373 and 0.6532 respectively on Flickr8k dataset. In the paper [21], authors investigated region based image captioning method along with knowledge graph along with encoder-decode of model to validate the generated captions. They evaluated the proposed work on MS-COCO and Flickr30k dataset. A hierarchical deep neural network is proposed for automatic image captioning in the paper [44]. The experimental results are evaluated on MS-COCO dataset.In the paper [7], authors proposed Bag-LSTM methods for automatic image captioning on MS-COCO dataset.Also, they proposed variants of LSTM on mentioned dataset and concluded that Bag-LSTM perform better on CIDEr value. In the paper [17], authors proposed fusion based text feature extraction for image captioning using DNN(deep neural network) with LSTM.They evaluated the proposed model on Fliker30k dataset. A semantic embedding as global guidance and attention model have proposed in the paper [22].The Experiments were conducted on Flickr8k, Flickr30k and MS-COCO to validate the proposed work. A R-CNN base top-down and bottom-up approach is proposed in paper [6]. The model is improved by re-ranked caption using beam search decoders and explanatory features.A Reference based Long Short Term Memory (R-LSTM) based method is proposed in the paper [11], for automatic image caption generation. They used weighting scheme between words and image to define relevant caption. Validation of proposed model was done on Fliker30k and MS-COCO dataset, and they commented on their experiment that CIDr value on MS-COCO dataset was increased 10.37%.

After going through the studies on image caption generation, mentioned above, some significant research points were identified. Firstly (RP1), CNN models used in most state-of-the-arts are pre-trained on ImageNet dataset which is object specific and not scene specific. Consequently, those models produce object specific results. Secondly (RP2), most papers reported their results in terms of one or two evaluation metrics only, such as accuracy and BLEU-1 score. To examine RP1, **VGG16_Hybrid_Places1365** model is used as CNN in the proposed model to provide object and scene specific result. This model has been pre-trained on both, the ImageNet and Places datasets. Furthermore, in order to address RP2, we evaluated the results using multiple evaluation metrics such as BLEU, METEOR, ROUGE and GLEU measures.

3 Methodology

This section discusses the methodology that has been employed in the present study. The main objective of the present study is to produce well-formed captions for input images. In this regard,

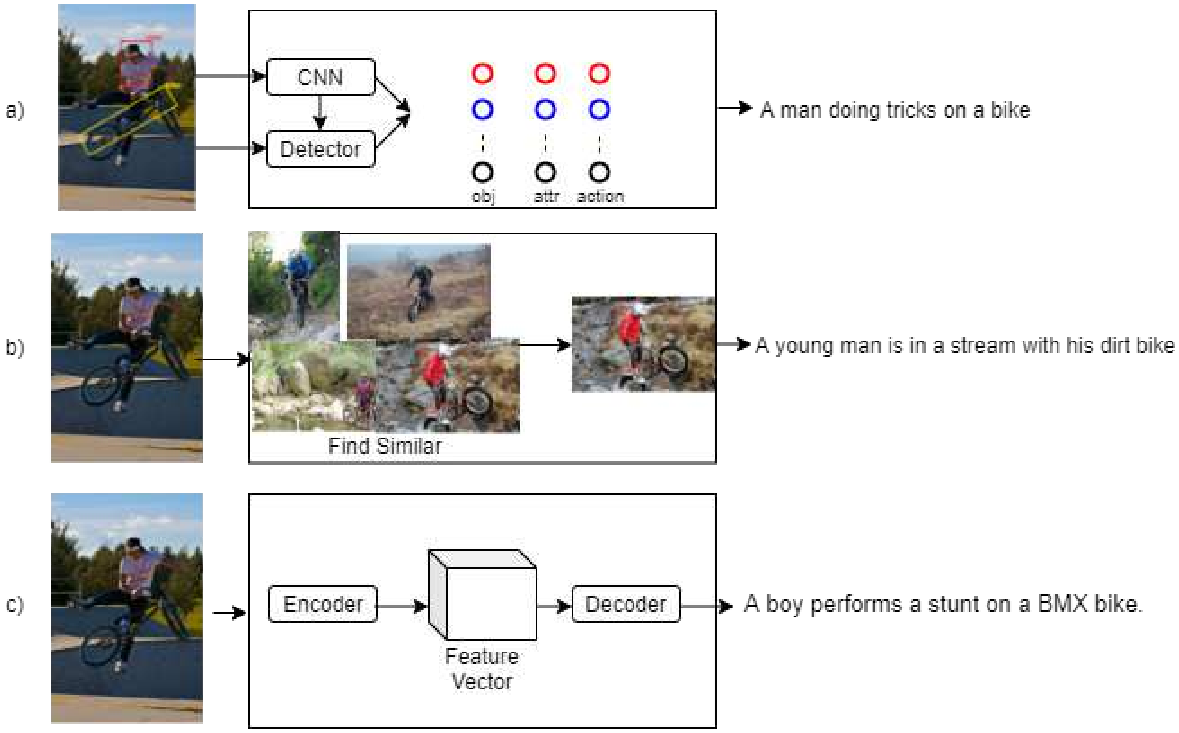

a)
CNN
Detector
obj
attr
action
A man doing tricks on a bike
b)
Find Similar
A young man is in a stream with his dirt bike
c)
Encoder
Feature Vector
Decoder
A boy performs a stunt on a BMX bike.

Fig. 1 Classification of image captioning models. (a) depicts template based approach to image captioning, (b) shows retrieval based captioning approach, (c) shows encoder-decoder based approach.

the following important ideas and concepts are relevant.

3.1 Feature Extraction

The encoder is used to extract an image's visual features. Convolutional neural networks (CNNs) are generally used as encoders. They are widely used as models for visual recognition tasks. Convolution layer, Pooling layer, and Fully connected layer are the three main layers that are present in all CNNs [2]. There are several pre-trained CNN-based models that are available in order to reduce the training time of the model. In this sub-section, some of the pre-trained CNN architectures used in previous research are discussed.

The first of these is 16-layered VGG16 model [53] that achieved 92.7% accuracy on ImageNet challenge. This model provides the features of images in the form of 1-D array that are used in caption generation. The main issue with VGG16 is its slow training if trained from the scratch. The second network is Inception V3 [47] model that achieves 4.2% error rate, which is lower than previously reported VGG16 results. This network requires less computational resources as compared to the VGG Net. Third, and the most recently developed model is ResNet50 [23], that is used for better learning in deep neural networks as compared to VGG16 and Inception V3. The parameters and number of layers of the respective models are shown in Table 1.

Table 1 Layers and parameters in CNN models.

S.No.	CNN	No. of Layers	No. of Parameters
1	VGG16	16	138,357,544
2	Inception V3	42	23,851,784
3	ResNet50	50	23,587,712

3.2 Sentence generation

The sentence generation component uses RNN based model to generate sentences. It is connected to the output of the feature extraction model. The basic RNN cannot handle long sequence of words efficiently. To solve these issues, the Long Short-Term Memory (LSTM) [19] network is employed.

The memory cell is the most important component of the LSTM architecture.

After successful use of LSTM for machine translation, sequence learning and speech recognition, it was found useful in image captioning [1] as well. Through memory cell and gates, LSTM can avoid vanishing and exploding gradient issues. Figure 2 shows three gates (input, output, and forget gate) controlled by reading and writing memory cell C. At time step i, LSTM receives inputs from various sources: X_i represents current input, H_{i-1} represents past hidden state, and C_{i-1} represents previous memory cell state. At time step t, the updated gate values for provided inputs X_i, H_{i-1}, and C_{i-1} are as follows:

$$I_i = \sigma(WM_{X_I}X_i + WM_{HI}h_{i-1} + BV_I) \tag{1}$$

$$F_i = \sigma(WM_{X_F}X_i + WM_{HF}h_{i-1} + BV_F) \tag{2}$$

$$O_i = \sigma(WM_{X_O}X_i + WM_{HO}h_{i-1} + BV_O) \tag{3}$$

$$G_i = \phi(WM_{X_C}X_i + WM_{HC}h_{i-1} + BV_C) \tag{4}$$

$$C_i = F_i * C_{i-1} + I_i * G_i \tag{5}$$

$$h_i = O_i * \phi(C_i) \tag{6}$$

Where X_I, X_F, X_O, X_C represent the inputs of the input gate, forget gate, output gate, and memory cell, WM represents weight metrics, and BV represents bias vectors.
The sigmoid activation function is ρ and given as below.

$$\rho(X) = \frac{1}{1 + expo(-X)}$$

ϕ is the hyperbolic tangent, given as below.

$$\phi(X) = \frac{expo(X) - expo(-X)}{expo(X) + expo(-X)}$$

$*$ is the product with the gate value.

3.3 Evaluation metrics

In order to measure the performance of the model, the evaluation process verifies the different characteristics of the generated captions such as readability, grammatical correctness and content of the caption. The assessment process of most of the metrics heavily depend on text matching between reference caption and the generated caption. In the current study, all popular evaluation metrics like BLEU [29], GLEU [38], ROUGE_L [48] and METEOR [31] are used to evaluate the proposed model.

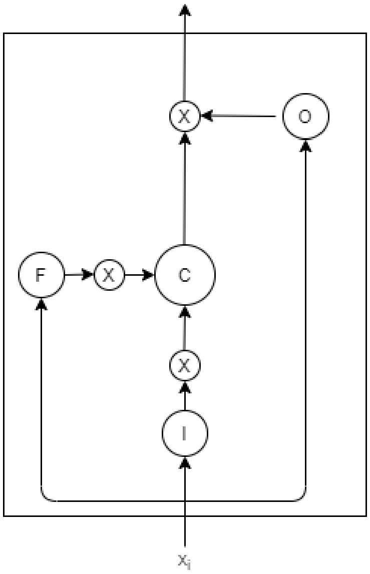


Fig. 2 Basic architecture of LSTM(Long Short Term Memory) cell.

The BLEU score is calculated by comparing the generated text's n-grams to the reference text's n-grams and counting the number of matches. It is a precision-based metric that is primarily intended for use in evaluation of machine translation results. In general, a BLEU score of greater than 0.5 is considered to be excellent while a score of less than 0.15 indicates that the model requires considerable improvements. One known issue with BLEU score is that it does not consider semantic meaning and structure of the sentences.

METEOR was created to unambiguously address the flaws in BLEU. METEOR score is computed using unigram-recall and unigram-precision. The METEOR metric relates more closely to the human perceived judgement as it gives more weightage to recall than precision as is also conjectured about human judgement.

The GLEU metric was created to assess sentence fluency and grammatical errors. It is estimated by overlapping n-grams with a set of reference sentences. It is incompatible with error precision/recall measures. On a corpus level, the GLEU score is quite similar to the BLEU score.

ROUGE_L is mostly used to assess image captioning quality. Using the longest common subsequence (LCS), it determines the words that have the longest matching sequence. It is not mandatory to make continuous word matches in this metric.

3.4 Dataset used

This study has used the popular Flickr8k [42] dataset for training and performance evaluation. This dataset contains manually labelled captions for each image. The dataset contains images and the corresponding captions in English language. The dataset has two parts: image directory and description file. The image directory contains 8000 images and for each image, there are 5 captions stored in the description file. Out of the total 8000 images, 6000 images are used for training, 1000 images were used for development and the rest 1000 images were used for testing purpose. A few sample images from the dataset, along with their reference captions in English, are shown in Figures 3 and 6. All the images are in jpg format. The average length of the captions is 12. The resolution of the input images is 256x500 to 500x500.

Furthermore, the MS COCO Captions dataset has been used for evaluation of the proposed model. This dataset contains over 164K images in everyday contexts. The objects in the images are classified into 80 classes. With respect to caption generation task, each image has 5 associated captions. The dataset is divided into training set (approx. 83K images), validation set (approx. 41K images) and test set (approx. 40K images).

3.5 Proposed Model

In this sub-section, a deep neural network based method of image caption generation is proposed. In this approach, an encoder-decoder based deep neural network model is designed with the help of convolutional neural network (VGG16 Hybrid Places1365) and recurrent neural network (LSTM) as decoder to generate captions for query images as shown in Figure 4. The VGG16 Hybrid Places1365 is used to generate the 1-D array representation of an image. Previously used CNN models such as VGG16, Inception, and ResNet are pre-trained on the 1000-classes ImageNet dataset, whereas the VGG16 Hybrid Places1365 model is trained on both ImageNet and Places datasets (containing 1000 and 365 classes respectively).

Given image I, the captured visual features are denoted as a vector $V = \{v_1, v_2, ...v_n\}$, which is calculated as:

$$V = CNN_c(I)$$

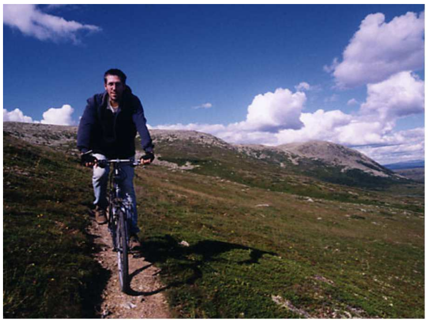
where $CNN_c(I)$ is the output of the last convolutional layer of the modified model.

The popular deep learning framework Tensorflow is used to design the model. Using Tensorflow, a CNN is developed as an encoder to extract visual data, and the extracted data is used as the LSTM decoder's initial state. The LSTM part of the model then generates the caption for the input image. The captions are generated one word at a time. Finally, the individual words are concatenated to produce a caption in sentence form.

After successful training of the proposed model, for a given input image I, caption is predicted one word at a time using the trained model. The probability of $P(S/I)$ [49] is maximized where S is the generated caption. The generated caption S which contains n words, creates a sequence $s_1, s_2, ..., s_n$ where each word is an element of vocabulary V. During the training, to generate the caption of images, a search method called Beam search is used. In beam search, the best n sentences are selected as a set of length $(t + 1)$ at a time t for evaluation. In this work, the model is trained and tested using a set of 5 reference captions, i.e., $(n = 5)$. The proposed model is trained using the Google Colab platform. The model is run for 10 epochs where each epoch took approximately 30 minutes to complete. The batch size used for training purpose is 256. The dropout value is taken to be 0.5. The activation function used in the layers of the model is ReLU while the 'adam' optimizer is used to minimize the 'categorical_crossentropy' loss.

4 Results and Analysis

There are two desirable characteristics expected from the generated caption after the model has been trained. Firstly, it should correlate with all the objects present in the image. Secondly, it should be useful and understandable to human beings.



1. A guy is riding a bike up the side of a hill.
2. A young man bicycles towards the camera and away from beautiful mountains on a clear day.
3. Man on bike in mountains.
4. Man riding a bicycle down a narrow path.
5. Man riding bike on trail.

Fig. 3 Sample image with reference captions.

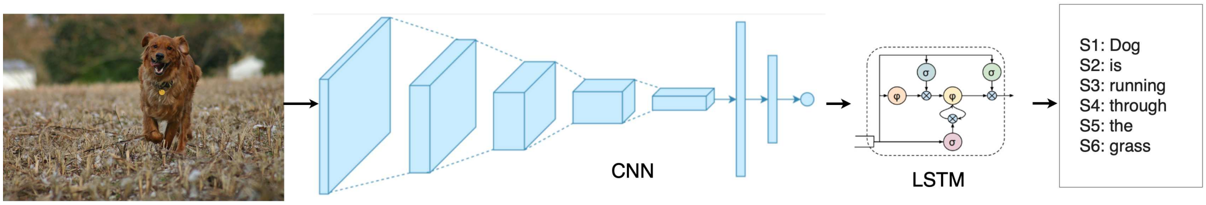


Fig. 4 Architecture of proposed model.

On the Flickr8k dataset, the proposed model yielded a BLEU score of 0.6666 or 66.66%. The following Table 2 shows the BLEU scores obtained on the Flickr8k dataset. Other applicable metrics such as ROUGE_L, GLEU and METEOR were determined after BLEU score had been calculated. Since precision, recall and F-Mean are necessary to compute ROUGE_L, they were obtained for the generated captions. The following Table 3 shows the ROUGE_L, GLEU and METEOR scores on Flickr8k dataset.

As shown in Tables 2 and 3, we experimented with a total of 10 different models. The study started with the Inception V3 with LSTM that yielded BLEU-1 score of 0.64. After refinement of this baseline model, we finally developed the VGG16 Hybrid Places 1365 model with LSTM. This model performed with a BLEU score of 0.6666 that was the best performance among all the experimented models. The same trend was observed with other popular metrics such as ROUGE_L, METEOR and GLEU with values 0.3076, 0.5060 and 0.2469 respectively.

While the proposed model provides comparatively better results, the following limitations were observed during the experimentation. Firstly, the generated captions were not following all the rules of English grammar. Secondly, high training time was observed and therefore, there is some scope for improving the efficiency of the training process. Last, but not the least, the generated captions appear to be less descriptive in comparison to the reference captions. Also, a few failure cases were noted during the experiments. For example, in Figure 6(j) , the image reference caption describes a woman in a pink shirt whereas the proposed model generates the caption “Man in red shirt is standing on the street.”. Again, for Figure 6(f), the reference caption describes a boy wearing red swim trunks while the proposed model outputs the caption “A child in red dress playing on the ground”.

4.1 Comparison with state-of-the-art

To show the efficacy of the proposed model, we compared the results obtained with the state-of-the-art models as shown in Tables 4 and 5 along with Figure 5. It can be clearly seen from the table that the proposed model outperformed all state-of-the-art approaches on Flickr8k and MSCOCO Captions datasets as measured by BLEU-1, BLEU-3 and BLEU-4 scores. The model is also comparable to the state of art performance according to BLEU-2 score. This indicates that the proposed model is capable to generate

Table 2 BLEU scores of proposed models on Flickr8k dataset. Emphasis indicates result of best model

S.No.	Encoder	Decoder	BLEU-1	BLEU-2	BLEU-3	BLEU-4
1	Inception V3	LSTM	0.6400	0.4131	0.2809	0.1571
2	Inception V3	B-LSTM	0.6121	0.3733	0.2465	0.3444
3	VGG16	LSTM	0.6400	0.3771	0.2985	0.1501
4	VGG16	B-LSTM	0.6274	0.3873	0.2670	0.1501
5	VGG16 Places365	LSTM	0.6382	0.4088	0.3040	0.2028
6	VGG16 Places365	B-LSTM	0.5956	0.2384	0.1484	0.2345
7	**VGG16 Hybrid Places1365**	LSTM	**0.6666**	**0.4340**	**0.3893**	**0.2878**
8	VGG16 Hybrid Places1365	B-LSTM	0.6441	0.3136	0.1760	0.2682
9	ResNet50	LSTM	0.5543	0.2863	0.1628	0.0877
10	VGG19	LSTM	0.5912	0.3321	0.2527	0.1732

Table 3 ROUGE_L, METEOR and GLEU scores of proposed models on Flickr8k dataset. Emphasis indicates result of best model.

S.No.	Encoder	Decoder	ROUGE-L	METEOR	GLEU
1	Inception V3	LSTM	0.1923	0.4	0.2157
2	Inception V3	B-LSTM	0.1199	0.4	0.1809
3	VGG16	LSTM	0.2264	0.2083	0.2134
4	VGG16	B-LSTM	0.1818	0.1626	0.1855
5	VGG16 Places365	LSTM	0.2592	0.4	0.1809
6	VGG16 Places365	B-LSTM	0.1923	0.4130	0.1776
7	**VGG16 Hybrid Places1365**	**LSTM**	**0.3076**	**0.5060**	**0.2469**
8	VGG16 Hybrid Places1365	B-LSTM	0.2692	0.4862	0.2157
9	ResNet50	LSTM	0.2541	0.4996	0.1312
10	VGG19	LSTM	0.1921	0.2000	0.1512

useful and human understandable captions. The effectiveness of the model is also confirmed by the results as obtained on other metric such as METEOR (0.5060 & 0.4768) and GLEU (0.2469 & 0.2798). It is important to note here that most of the state-of-the-art works do not report their results on these mentioned metrics.

4.2 Sample result caption

In order to demonstrate the validity of the proposed model, we show the generated captions for ten sample images from the dataset along with their reference captions (ground-truth) in Figure 6. The result on these random images further strengthens the claimed effectiveness of the model.

5 Conclusion and Future work

This study proposed an encoder-decoder based model to generate grammatically correct captions for images. The proposed model consists of a CNN-based encoder and an LSTM-based decoder. The model has been evaluated on various standard metrics such as BLEU, METEOR, GLEU and ROUGE_L with Flickr8k and MSCOCO Captions datasets. The experimental results show that the model surpasses all existing state-of-the-art approaches in BLEU, METEOR and GLEU scores. The paper also reported results of caption generation on live sample images that reinforces the validity of the proposed approach.

As the proposed model was very close to the best results as per ROUGE_L score, the future approaches may improve on this aspect.

Table 4 Performance comparison of proposed model against state-of-the-art techniques on Flickr8K dataset (- indicates that the result is not reported in the paper, B-1: BLEU-1, B-2: BLEU-2, B-3: BLEU-3, B-4: BLEU-4, RL: ROUGE_L, MT: METEOR, GL: GLEU)

S.No.	Model	B-1	B-2	B-3	B-4	RL	MT	GL
1	M-RNN (2014) [34]	0.5778	0.2751	0.2307	-	-	-	-
2	NIC (2015) [49]	0.6300	0.4100	0.2700	-	-	-	-
3	M-RNN (2015) [33]	0.5650	0.3860	0.2560	0.1700	-	-	-
4	V-S.M(2015) [24]	0.579	0.383	0.245	0.160	-	-	-
5	DL(2018) [3]	0.5335	-	-	-	-	-	-
6	NNM(2019) [15]	0.5600	-	-	-	-	-	-
7	AICRL (2020) [10]	0.619	0.452	0.3668	0.262	-	0.209	-
8	EL(2020) [25]	0.634	0.400	0.287	0.151	-	-	-
9	CL (2021) [45]	0.6373	0.4500	0.3087	0.2113	0.4641	0.1995	-
10	CLA (2021) [45]	0.6532	0.4692	0.3281	0.2258	0.4695	0.2087	-
11	**Proposed Model**	**0.6666**	**0.4704**	**0.3893**	**0.2878**	**0.3076**	**0.5060**	**0.2469**

Table 5 Performance comparison of proposed model against state-of-the-art techniques on MSCOCO Captions dataset (- indicates that the result is not reported in the paper, B-1: BLEU-1, B-2: BLEU-2, B-3: BLEU-3, B-4: BLEU-4, RL: ROUGE_L, MT: METEOR, GL: GLEU)

S.No.	Model	B-1	B-2	B-3	B-4	RL	MT	GL
1	NIC (2015) [49]	0.666	0.461	0.329	0.246	-	-	-
2	M-RNN (2015) [33]	0.67	0.49	0.35	0.25	-	-	-
3	V-S.M(2015) [24]	0.625	0.45	0.321	0.23	-	0.195	-
4	DL(2018) [3]	0.67257	-	-	-	-	-	-
5	AICRL (2020) [10]	0.731	0.562	0.41	0.326	-	0.261	-
6	**Proposed Model**	**0.7350**	**0.5421**	**0.4233**	**0.3342**	**0.3566**	**0.4768**	**0.2798**

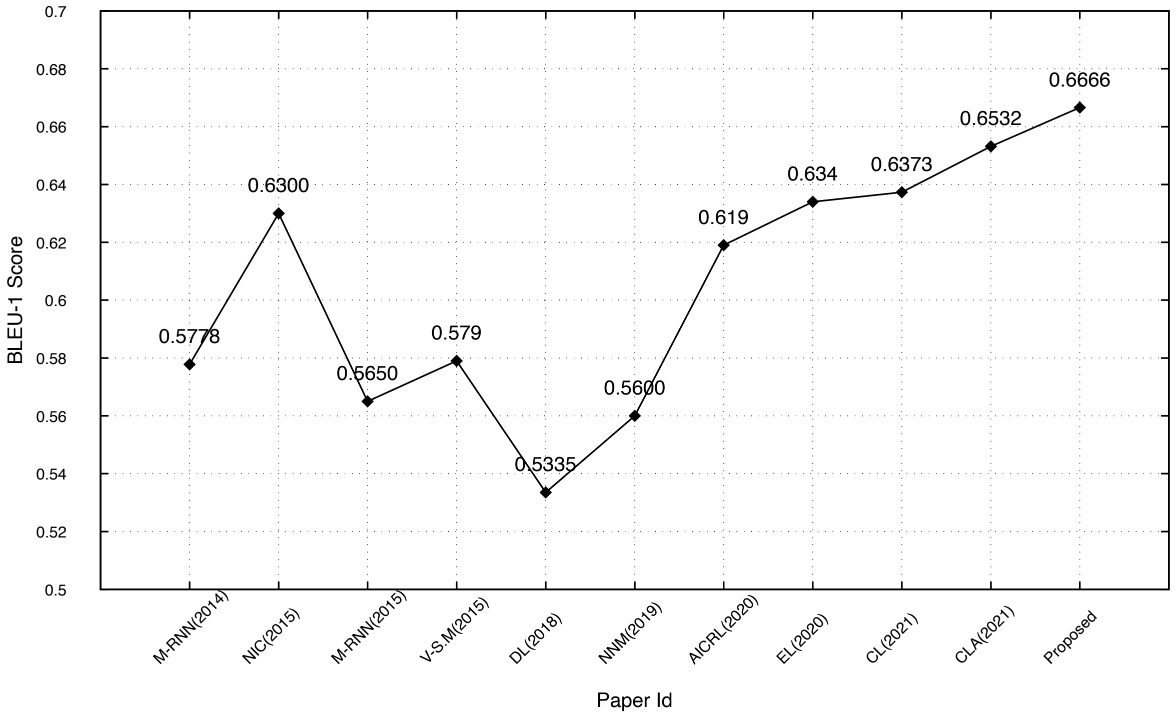
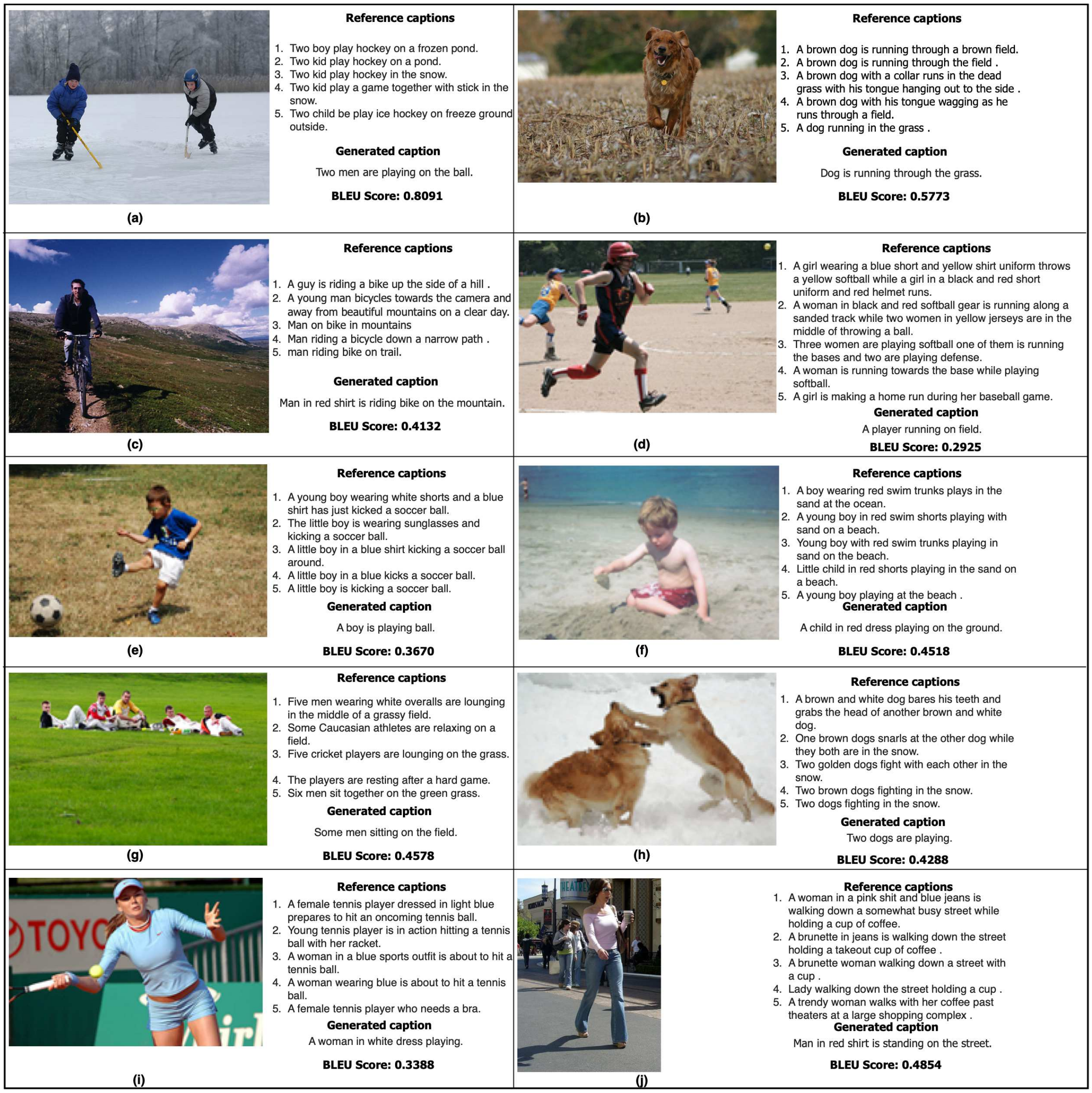


Fig. 5 Graph comparison of BLEU-1 scores of proposed model against state-of-the-art techniques on Flickr8k dataset.



Reference captions

1. Two boy play hockey on a frozen pond.
2. Two kid play hockey on a pond.
3. Two kid play hockey in the snow.
4. Two kid play a game together with stick in the snow.
5. Two child be play ice hockey on freeze ground outside.

Generated caption

Two men are playing on the ball.

BLEU Score: 0.8091

(a)

Reference captions

1. A brown dog is running through a brown field.
2. A brown dog is running through the field .
3. A brown dog with a collar runs in the dead grass with his tongue hanging out to the side .
4. A brown dog with his tongue wagging as he runs through a field.
5. A dog running in the grass .

Generated caption

Dog is running through the grass.

BLEU Score: 0.5773

(b)

Reference captions

1. A guy is riding a bike up the side of a hill .
2. A young man bicycles towards the camera and away from beautiful mountains on a clear day.
3. Man on bike in mountains
4. Man riding a bicycle down a narrow path .
5. man riding bike on trail.

Generated caption

Man in red shirt is riding bike on the mountain.

BLEU Score: 0.4132

(c)

Reference captions

1. A girl wearing a blue short and yellow shirt uniform throws a yellow softball while a girl in a black and red short uniform and red helmet runs.
2. A woman in black and red softball gear is running along a sanded track while two women in yellow jerseys are in the middle of throwing a ball.
3. Three women are playing softball one of them is running the bases and two are playing defense.
4. A woman is running towards the base while playing softball.
5. A girl is making a home run during her baseball game.

Generated caption

A player running on field.

BLEU Score: 0.2925

(d)

Reference captions

1. A young boy wearing white shorts and a blue shirt has just kicked a soccer ball.
2. The little boy is wearing sunglasses and kicking a soccer ball.
3. A little boy in a blue shirt kicking a soccer ball around.
4. A little boy in a blue kicks a soccer ball.
5. A little boy is kicking a soccer ball.

Generated caption

A boy is playing ball.

BLEU Score: 0.3670

(e)

Reference captions

1. A boy wearing red swim trunks plays in the sand at the ocean.
2. A young boy in red swim shorts playing with sand on a beach.
3. Young boy with red swim trunks playing in sand on the beach.
4. Little child in red shorts playing in the sand on a beach.
5. A young boy playing at the beach .

Generated caption

A child in red dress playing on the ground.

BLEU Score: 0.4518

(f)

Reference captions

1. Five men wearing white overalls are lounging in the middle of a grassy field.
2. Some Caucasian athletes are relaxing on a field.
3. Five cricket players are lounging on the grass.
4. The players are resting after a hard game.
5. Six men sit together on the green grass.

Generated caption

Some men sitting on the field.

BLEU Score: 0.4578

(g)

Reference captions

1. A brown and white dog bares his teeth and grabs the head of another brown and white dog.
2. One brown dogs snarls at the other dog while they both are in the snow.
3. Two golden dogs fight with each other in the snow.
4. Two brown dogs fighting in the snow.
5. Two dogs fighting in the snow.

Generated caption

Two dogs are playing.

BLEU Score: 0.4288

(h)

Reference captions

1. A female tennis player dressed in light blue prepares to hit an oncoming tennis ball.
2. Young tennis player is in action hitting a tennis ball with her racket.
3. A woman in a blue sports outfit is about to hit a tennis ball.
4. A woman wearing blue is about to hit a tennis ball.
5. A female tennis player who needs a bra.

Generated caption

A woman in white dress playing.

BLEU Score: 0.3388

(i)

Reference captions

1. A woman in a pink shit and blue jeans is walking down a somewhat busy street while holding a cup of coffee.
2. A brunette in jeans is walking down the street holding a takeout cup of coffee .
3. A brunette woman walking down a street with a cup .
4. Lady walking down the street holding a cup .
5. A trendy woman walks with her coffee past theaters at a large shopping complex .

Generated caption

Man in red shirt is standing on the street.

BLEU Score: 0.4854

(j)

Fig. 6 Figure showing sample images with their reference captions along with generated caption using proposed model.

Also, other available metrics may also be used to report experimental results. Further, attention-based models may be employed for strengthening the robustness of the models [26]. Also, other cross-domain approaches such as neuro-symbolic approaches may be preferred in order to enhance the explain ability of proposed models along with the reasoning for generating specific captions for corresponding images [41]. Image captioning may also be extended to video inputs to generate captions for detecting important events in video over a time range. This may be highly relevant for applications such as surveillance, home monitoring, etc.

Declarations

- Funding: No funds, grants, or other support was received.
- Competing interests: The authors have no competing interests to declare that are relevant to the content of this article.

- Ethics approval: Not applicable
- Consent to participate: Not applicable
- Consent for publication: All authors agree to the publication of the final manuscript.
- Availability of data and materials: Not applicable
- Code availability: The code will be made available if required.
- Authors' contributions: Arun Kumar Yadav, Mohit Kumar and Divakar Yadav contributed to the study conception and design. Material preparation, data collection and analysis were performed by Akash Verma. The first draft of the manuscript was written by Akash Verma and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

References

[1] A. Graves, A. Mohamed and G. E. Hinton. Speech recognition with deep recurrent neural networks. pages 6645–6649, 2013.

[2] Saad Albawi and Tareq Abed Mohammed. Understanding of a Convolutional Neural Network. 2017.

[3] Chetan Amritkar and Vaishali Jabade. Image Caption Generation Using Deep Learning Technique. *Proceedings - 2018 4th International Conference on Computing, Communication Control and Automation, ICCUBEA 2018*, pages 1–4, 2018.

[4] Georgios Barlas, Christos Veinidis, and Avi Arampatzis. What we see in a photograph: content selection for image captioning. *The Visual Computer*, 37(6):1309–1326, 2021.

[5] Khaled Bayoudh, Raja Knani, Fayçal Hamdaoui, and Abdellatif Mtibaa. A survey on deep multimodal learning for computer vision: advances, trends, applications, and datasets. *The Visual Computer*, pages 1–32, 2021.

[6] Rajarshi Biswas, Michael Barz, and Daniel Sonntag. Towards explanatory interactive image captioning using top-down and bottom-up features, beam search and re-ranking. *KI-Künstliche Intelligenz*, 34(4):571–584, 2020.

[7] Pengfei Cao, Zhongyi Yang, Liang Sun, Yanchun Liang, Mary Qu Yang, and Renchu Guan. Image captioning with bidirectional semantic attention-based guiding of long short-term memory. *Neural Processing Letters*, 50(1):103–119, 2019.

[8] Hui Chen, Guiguang Ding, Zijia Lin, Yuchen Guo, Caifeng Shan, and Jungong Han. Image captioning with memorized knowledge. *Cognitive Computation*, 13(4):807–820, 2021.

[9] Yejin Choi, Tamara L Berg, U N C Chapel Hill, Chapel Hill, and Stony Brook. T REE T ALK : Composition and Compression of Trees for Image Descriptions. 2:351–362, 2014.

[10] Yan Chu, Xiao Yue, Lei Yu, Mikhailov Sergei, and Zhengkui Wang. Automatic image captioning based on resnet50 and lstm with soft attention. *Wireless Communications and Mobile Computing*, 2020, 2020.

[11] Guiguang Ding, Minghai Chen, Sicheng Zhao, Hui Chen, Jungong Han, and Qiang Liu. Neural image caption generation with weighted training and reference. *Cognitive Computation*, 11(6):763–777, 2019.

[12] Jeff Donahue, Lisa Anne Hendricks, Marcus Rohrbach, Subhashini Venugopalan, Sergio Guadarrama, Kate Saenko, and Trevor Darrell. Long-Term Recurrent Convolutional Networks for Visual Recognition and Description. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(4):677–691, 2017.

[13] Xinzhi Dong, Chengjiang Long, Wenju Xu, and Chunxia Xiao. Dual graph convolutional networks with transformer and curriculum learning for image captioning. *arXiv preprint arXiv:2108.02366*, 2021.

[14] Ali Farhadi, Mohsen Hejrati, Mohammad Amin Sadeghi, Peter Young, Cyrus Rashtchian, Julia Hockenmaier, and David Forsyth. Every picture tells a story: Generating sentences from images. In *European conference on computer vision*, pages 15–29. Springer, 2010.

[15] Ayan Ghosh, Debarati Dutta, and Tiyasa Moitra. A Neural Network Framework to Generate Caption from Images. *Springer Nature Singapore Pte Ltd.*, pages 171–180, 2020.

[16] Yunchao Gong, Liwei Wang, Micah Hodosh, Julia Hockenmaier, and Svetlana Lazebnik. Improving image-sentence embeddings using large weakly annotated photo collections.

Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 8692 LNCS(PART 4):529–545, 2014.

[17] Neeraj Gupta and Anand Singh Jalal. Integration of textual cues for fine-grained image captioning using deep cnn and lstm. *Neural Computing and Applications*, 32(24):17899–17908, 2020.

[18] Chen He and Haifeng Hu. Image captioning with text-based visual attention. *Neural Processing Letters*, 49(1):177–185, 2019.

[19] Sepp Hochreiter and Jürgen Schmidhuber. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780, 1997.

[20] Micah Hodosh, Peter Young, and Julia Hockenmaier. Framing image description as a ranking task: Data, models and evaluation metrics. *IJCAI International Joint Conference on Artificial Intelligence*, 2015-Janua(Ijcai):4188–4192, 2015.

[21] Feicheng Huang, Zhixin Li, Haiyang Wei, Canlong Zhang, and Huifang Ma. Boost image captioning with knowledge reasoning. *Machine Learning*, 109(12):2313–2332, 2020.

[22] Teng Jiang, Zehan Zhang, and Yupu Yang. Modeling coverage with semantic embedding for image caption generation. *The Visual Computer*, 35(11):1655–1665, 2019.

[23] Jian Sun Kaiming He, Xiangyu Zhang, Shaoqing Ren. Deep Residual Learning for Image Recognition. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[24] Andrej Karpathy and Li Fei-Fei. Deep Visual-Semantic Alignments for Generating Image Descriptions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(4):664–676, 2017.

[25] Harshitha Katpally and Ajay Bansal. Ensemble learning on deep neural networks for image caption generation. *Proceedings - 14th IEEE International Conference on Semantic Computing, ICSC 2020*, pages 61–68, 2020.

[26] Muhammad Jaleed Khan and Edward Curry. Neuro-symbolic visual reasoning for multimedia event processing: Overview, prospects and challenges. In *CIKM (Workshops)*, 2020.

[27] Muhammad Junaid Khan, Muhammad Jaleed Khan, Adil Masood Siddiqui, and Khurram Khurshid. An automated and efficient convolutional architecture for disguise-invariant face recognition using noise-based data augmentation and deep transfer learning. *The Visual Computer*, pages 1–15, 2021.

[28] Ryan Kiros, Ruslan Salakhutdinov, and Richard Zemel. Multimodal neural language models. *31st International Conference on Machine Learning, ICML 2014*, 3:2012–2025, 2014.

[29] Todd Ward Kishore Papineni, Salim Roukos and WeiJing Zhu. BLEU: a Method for Automatic Evaluation of Machine Translation. *Annalen der Physik*, 371(23):437–461, 1922.

[30] Girish Kulkarni, Visruth Premraj, Vicente Ordonez, Sagnik Dhar, Siming Li, Yejin Choi, Alexander C. Berg, and Tamara L. Berg. Baby talk: Understanding and generating simple image descriptions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(12):2891–2903, 2013.

[31] Alon Lavie and Abhaya Agarwal. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. *Proceedings of the Second Workshop on Statistical Machine Translation*, (June):228–23, 2007.

[32] Xiaoxiao Liu, Qingyang Xu, and Ning Wang. A survey on deep neural network-based image captioning. *The Visual Computer*, 35(3):445–470, 2019.

[33] Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, Zhiheng Huang, and Alan Yuille. Deep captioning with multimodal recurrent neural networks (m-rnn). *arXiv preprint arXiv:1412.6632*, 2014.

[34] Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, and Alan L Yuille. Explain images with multimodal recurrent neural networks. *arXiv preprint arXiv:1410.1090*, 2014.

[35] Rebecca Mason and Eugene Charniak. Nonparametric Method for Data-driven Image Captioning. pages 592–598, 2014.

[36] Md. Zakir Hossain, Ferdous Sohel, Mohd Fairuz Shiratuddin and Hamid Laga. A Comprehensive Survey of Deep Learning for Image Captioning. *ACM Comput. Surv. Articl*, 0(0):36, 2018.

[37] Margaret Mitchell, Xufeng Han, Jesse Dodge, Alyssa Mensch, Amit Goyal, Alex Berg,

Kota Yamaguchi, Tamara Berg, Karl Stratos, and Hal Daumé. Midge: Generating image descriptions from computer vision detections. *EACL 2012 - 13th Conference of the European Chapter of the Association for Computational Linguistics, Proceedings*, pages 747–756, 2012.

[38] Andrew Mutton, Mark Dras, Stephen Wan, and Robert Dale. GLEU : Automatic Evaluation of Sentence-Level Fluency. (June):344–351, 2007.

[39] Vicente Ordonez, Xufeng Han, Polina Kuznetsova, Girish Kulkarni, Margaret Mitchell, Kota Yamaguchi, Karl Stratos, Amit Goyal, Jesse Dodge, Alyssa Mensch, et al. Large scale retrieval and generation of image descriptions. *International Journal of Computer Vision*, 119(1):46–59, 2016.

[40] Vicente Ordonez, Girish Kulkarni, and Tamara L. Berg. Im2Text: Describing images using 1 million captioned photographs. *Advances in Neural Information Processing Systems 24: 25th Annual Conference on Neural Information Processing Systems 2011, NIPS 2011*, pages 1–9, 2011.

[41] Yuqing Peng, Chenxi Wang, Yixin Pei, and Yingjun Li. Video captioning with global and local text attention. *The Visual Computer*, pages 1–12, 2021.

[42] Cyrus Rashtchian, Peter Young, Micah Hodosh, and Julia Hockenmaier. Collecting image annotations using amazon's mechanical turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 139–147, 2010.

[43] Richard Socher, Andrej Karpathy, Quoc V. Le, Christopher D. Manning, and Andrew Y. Ng. Grounded Compositional Semantics for Finding and Describing Images with Sentences. *Transactions of the Association for Computational Linguistics*, 2:207–218, 2014.

[44] Yuting Su, Yuqian Li, Ning Xu, and An-An Liu. Hierarchical deep neural network for image captioning. *Neural Processing Letters*, 52(2):1057–1067, 2020.

[45] Department of Computer Science Sulabh Katiyar, Samir Kumar Borgohain and Silchar Engineering National Institute of Technology. Comparative evaluation of cnn architectures for image caption generation. *International Journal of Advanced Computer Science and Applications*, 2021.

[46] Chen Sun, Chuang Gan, and Ram Nevatia. Automatic concept discovery from parallel text and visual corpora. *Proceedings of the IEEE International Conference on Computer Vision*, 2015 Inter:2596–2604, 2015.

[47] Christian Szegedy, Vincent Vanhoucke, Jonathon Shlens, and Zbigniew Wojna. Rethinking the Inception Architecture for Computer Vision. *Conference on Computer Vision and Pattern Recognition*, pages 2818–2826, 2016.

[48] Goro Tsuchiya. Postmortem Angiographic Studies on the Intercoronary Arterial Anastomoses.: Report I. Studies on Intercoronary Arterial Anastomoses in Adult Human Hearts and the Influence on the Anastomoses of Strictures of the Coronary Arteries. *Japanese Circulation Journal*, 34(12):1213–1220, 1971.

[49] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 07-12-June:3156–3164, 2015.

[50] Fen Xiao, Xue Gong, Yiming Zhang, Yanqing Shen, Jun Li, and Xieping Gao. DAA: Dual LSTMs with adaptive attention for image captioning. *Neurocomputing*, 364:322–329, 2019.

[51] Yezhou Yang, Ching Lik Teo, Hal Daumé, and Yiannis Aloimonos. Corpus-guided sentence generation of natural images. *EMNLP 2011 - Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, (May 2014):444–454, 2011.

[52] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. Image captioning with semantic attention. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2016-Decem:4651–4659, 2016.

[53] Karen Simonyan & Andrew Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. pages 1–14, 2015.