Automatic Punjabi Caption Generation For Sports Images

Manleen Kaur[1]
Gurpreet Singh Josan[1]
Jagroop Kaur[2]

[1] Department of Computer Science
Punjabi University Patiala
147002, Punjab India.
manleensaini@gmail.com,josangurpreet@pbi.ac.in

[2]Department of Computer Science and Engineering
Punjabi University Patiala
147002, Punjab India.
jagroop_80@rediffmail.com

Abstract. Image understanding and language generation have always been a difficult task in the field of Artificial Intelligence. Automatic Image Caption Generation is concerned with the task of understanding the image and generating a caption for it. In this paper, we represented our research work that uses the Deep Learning technique to create Punjabi captions for a given image and its associated news document. High-level features of the images are extracted using the pre-trained VGG-19 (Visual Geometry Group) model. These image features are merged with features of news text which are extracted using LSTM (Long Short Term Memory). The proposed model augments keywords from associated news text to generate suitable captions. Using both BLEU scores and human evaluations, we show that the proposed method is successful in generating intelligible and suitable captions.

Keywords: Image Caption, Deep Neural Network, Sequence to Sequence generation, Keyword Augmentation.

(Received May 19th, 2021 / Accepted June 1st, 2021)

1 Introduction

Among various tasks in the field of Artificial Intelligence, automatic caption generation is very challenging as it involves abstracting visual space and transforming that space to textual space. It includes an understanding of the image in terms of objects contained in it as well as a description of those understandings in the form of sentences. An image is incomplete until it gets any caption. The absence of a caption leads to speculation about the image where a person starts making his own assumptions of the image without knowing what the image is actually about. A caption conveys the information that describes the activities or the happenings in the image and who performs those activities. The caption is describing the image in a single line. Although a lot of research is focused on reducing the semantic gap between visual space and textual space with some success, still the results are partial and unstructured. Systems can find out different objects and actions in an image but relating those objects with appropriate actions is still a challenging problem.

Generating captions automatically can be helpful in various areas. Visual data is growing by

leaps and bounds and is available easily. This data contains much information. This raises an urgent demand for semantic understanding of images and representing that understanding in the textual form so that it can be made available through search engines. Organizing such data is the demand of the hour. It can be helpful in many domains which include the health sector, surveillance, entertainment, print media, etc. Various Information Retrieval Systems (IRS) indexed images based on the text associated with them. This is the oldest way of categorizing images in libraries. But images frequently have little or no accompanying textual information along with them. An automated caption generation system will help in better indexing of images and improving the efficiency of Information Retrieval Systems.

A caption is an important term for news articles also. A good caption represents the image and the accompanying document succinctly and makes the news attractive. Captions make the interest of reader into the article. Accomplishing this task could be advantageous as a news media management tool, for visually impaired people, advertisers, and corporate sectors, in military and multimedia applications. Many of the search engines retrieve the images without inspecting the image's content i.e., based on file format and file name, text surrounding the images, caption, etc. Thus the automatic generation of a caption for images helps in better search results and can help in indexing images so that exact information can be retrieved. Such a system can also be useful in suggesting captions to the users for their pictures which they are sharing on various social networking sites. Real-world is conceptualized by human beings using different modalities including five senses (touch, smell, vision, hearing, and taste) and background knowledge about the real world. Images provide only visual information. Human beings combined this information with their own world knowledge to generate captions. Thus caption is a highly abstractive view of visual representation.

Although captions generated by a human mind has no match, it has its own disadvantages. It is a very time-consuming process and depends on the background knowledge of the human being. On the other hand, automating the caption generation process is a potential alternative but not a trivial task. Limited word knowledge of the machine, its lack of ability to generate grammatically correct sentences, scarcity of resources are the major issues. The available dataset could be noisy. It may not provide a good caption and needs pre-processing first. The whole process is dependent on the object's recognition in the image. It is hard to extract features from low-resolution images making it hard to recognize the objects and their actions clearly, thus results in a bad caption. Yet another problem is a mismatch between images and their associated captions. The information contained in the image does not match with the caption defined against it. Captions based upon simple keywords are possible but abstractive captions based on the semantics of images are still a distant dream.

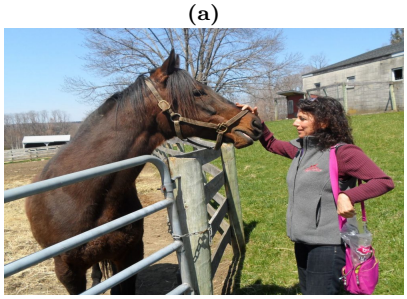
The field of automatic caption generation is a confluence of two different streams-computer vision and natural language processing. Computer vision involves the identification of objects in an image, their categories, activities, and other high-level features of images whereas natural language processing deals with text analysis and generation. Both fields have extensively studied in the past which made the foundation of research in exploiting the best of both worlds to automate the process of generating text from visual information. The major challenges in this area are:

- Identification of objects and their interaction: Computer vision techniques are mature enough to recognize all major objects in an image but still lacks in identifying objects appearing at distant or mixed with other objects. Some time some objects are blurred. Techniques are still developing to detect the interaction among objects. For example in figure 1a, it is hard for a system to recognize that shadow on the ground is not another player. Similarly, in figure 1b, it is hard for the system to recognize the action of a person as loving the horse or feeding the horse.

- Ranking detected objects in an image: Once objects are identified in an image, the system can not understand the importance of objects according to their relevance in the image. Accordingly, it can not mark more important objects from the scenario which should contribute more towards the generation of the textual representation of an image.

- Not able to recognize specific objects: Current systems are trained on general objects and can recognize only those objects for which they are trained. The system will not detect the name of the person or location from the image.

(a)



(b)

(c)

Figure 1: Ambiguous images for computer vision

For example in figure 1c, the system may find out objects like person, trophy, ground, crowd, etc., and able to generate general caption like "A person with a trophy". It is hard for the system to recognize that the person is "Virat Kohli" and the trophy is a "World Cup".

- Lack of standard dataset: Another concern is the lack of a standard dataset. A standard data set is required for automatic evaluation and comparing systems with each other.

Captions can be generated using extractive or abstractive techniques. The extractive approach focuses on sentence extraction based on the similarity of description keywords with the sentences in the document. These description keywords are the extracted features from the input image [8], [12]. There are three ways of finding the similarity viz. word overlap, cosine similarity, and probabilistic similarity. Abstractive captions are generated considering the semantics of images. Word-based Model and phrase-based model are two techniques of abstractive caption generation. The word-based model determines the probability of the word appearing in the caption given the image and its associated document. The words are the output of the content extraction of the input image. The phrase-based Model selects a subset from the set of phrases and combined them to compose the complete sentence. In the past few years, the deep neural network has been proved to be efficient for sequence generation and attracts the researcher's attention for generating caption of the image by using image features obtained through the deep neural network. Convolution Neural Network provides the rich representation of the input image and combining it with the Long Short Term Memory model for converting the image representation into description works as an icing on the cake. Convolution Neural Network (CNN) works as an encoder that encodes the information of the image and LSTM works as a decoder that decodes the information received from the CNN into a meaningful description. A supervised learning method has been used for caption generation where the system was trained using images having multiple captions. Datasets for the English language have been available. But for the other language where multiple captions are not available, the performance of the system degrades. The performance of the system can be improved by combining external information with images. News articles can be strong candidates for external information where the text of the article can provide contextual information of the image. This paper explains our idea of combining text features with image features for obtaining the caption of the image. A system has been proposed that extracts the higher-level features of the given image using a pre-trained VGG-19 (Visual Geometry Group) model and combines it with the features of the news text. The system is based on an encoder-decoder architecture. We experimented with various combinations for combining text and image features, attention mechanism. A further enhancement is observed by initializing the decoder with keywords extracted from the news articles. The contribution of this paper is:

- Proposes a multi-model architecture for generating more fluent and specific caption by integrating features from the image and associated text.

- Development of dataset for images and corre-

sponding Punjabi language captions

Empirical evaluations using both automatic metrics (BLEU Score) and human evaluation were performed, and it has been exhibited that the captions generated by the proposed approach were more fluent and correct than the baseline model. The next section discusses related work. Section 3 covers the proposed model architecture followed by experimental setup and results. Finally, we conclude the paper and outline future directions.

2 Related Work

One of the earlier methods given by [8] rely on the image annotation model (Scale Invariant Feature Transform Algorithm) to provide description keywords and based on these keywords two approaches (Extractive and Abstractive Caption Generation) are applied to generate a caption for image and its associated news. The extractive approach extracts the sentence from the news document based on the similarity of the keywords extracted. The abstractive approach generates the sentence based on the phrase and word-based model. [14], [11], [27] showed that image captioning is a solvable problem. Later [19] proposed a method that extracts the keywords from the image and based on those keywords, retrieved the similar images from the database and ranking is done on their respective captions, the best caption is selected. [31] utilized an annotated image dataset and employed a natural language generation (NLG) method that converts the image parsing results into textual descriptions. [15] proposed a method that takes an image as input. Phrases of similar images to the query image are retrieved from the database. Then combine those phrases to generate a description of a query image. [9] uses an LDA-based model for image annotation and use a wide variety of surface realization techniques to generate captions. Later, [18] proposed a method that takes a domain-specific query image and extracts the caption similar to the query image and performs sentence compression on it that reduces the length of the sentence without changing its meaning and deletes the modifiers so that the resulting caption is relevant to the query image. [5] proposed an approach that finds 'k' nearest images to the query image using cosine similarity and the image having the highest similarity score, its caption is considered as a final caption. These approaches generate a set of keywords and then tries to make phrases out of suggested keywords. Thus the systems may generate phrases that are semantically unrelated. These systems also have limited capability to generalize.

Deep Neural Network technique was first used by [28]. They presented a single joint model that uses a convolution neural network for image feature extraction. By embedding the input image into a fixed-length vector and input to the Recurrent Neural Network (RNN) is given by the last hidden layer that generates the sentence. [29] proposed an attention-based method which is modification on [28]. CNN extracts the features of the image and attention allows the most important features to dynamically come to the front. Then LSTM is used to produce the caption generated one word at each time step. [6] proposed a method that uses the VGG model for image feature extraction and further Principal Component Analysis (PCA) is used to reduce the dimension of an image feature, which is then provided as an input to LSTM that predicts one word at a time until the emission of the last word. [9], [29] are some other work in this area. In previous researches, there was a limitation of LSTM that each step does not have an association with the input image. Hence, only the first step receives the image. The information about the image fades away in successive steps as a result of gates. A comprehensive survey of all techniques is provided by [10], [16] and [22]. Similar to our work is presented by [2] and [34]. Both use external information to produce more fluent and correct captions. [33] gives an attention mechanism to detect the visual keywords more accurately using optimized pointer network where object detector detects the salient object level features which were not accurate in previous researches. [1] uses Region-CNN approach and sentiment analysis to extract the features of the image and output of which is passed to LSTM for caption generation. Further, adaptive sorting is done on the basis of Edge Rank algorithm with extra parameters. Feedback from humans is taken which works as a criterion to assign the edge weights for improving the chance of getting best caption in one go. [7] used similar approach for extracting the image features and Bidirectional Gated Recurrent Unit worked as a decoder to generate the captions in Bengali language. Argmax and Beam search is used to produce the quality captions. Similar to our work is presented by [2], [34] and [30]. They use external in-formation to produce more fluent and correct captions. [30] is the generalisation of [3]. They

used a multimodal transformer model approach by having an image encoder (CNN – that recognise object and scenes both), image-article encoder (encoder module and image attending module) and a decoder (multi-head target source attention). This model suffered from the low score of both in human evaluation and BLEU score which is 18.78. [34] utilizes entity labels produced by some upstream model as input to the captioning model and tries to produce fine-grained captions. [2] on the other hand utilizes news articles for extracting useful features from text combine them with image features to generate vectors. The most similar sentence was retrieved as a caption from the original news article based on cosine similarity. Effectively, captions are the sentences from news articles and not the exact description of image. Our approach is a multi-model approach that integrates both text and image features and uses keywords to generate captions.

3 Encoder Decoder Architecture

Encoder decoder architecture is a core to any sequence to sequence generation task. The encoder model is used to extract features of the source side and the decoder model takes encoded features and tries to produce a sequence based on these. The success of encoder-decoder architecture in machine translation tasks inspires its use in caption generation. For text data, the encoder consists of some variant of recurrent neural network (generally LSTM) whereas the convolution neural network is more suitable for extracting features of images. Several off the shelf pre-trained CNN networks are available to extract features of the image like AlexNet, GoogleNet, and VGGNet. The last hidden layer of these networks is treated as a vector. [4] found that the VGGNet model is better over the other models by comparing the three models over the BLEU score. Decoder generally consists of a bi-directional LSTM network. The attention mechanism is generally included which dynamically brings relevant features into consideration as required. Soft and hard attention mechanisms are proposed by [32]. The encoder combines the features of image, text, and caption and encodes in a fixed-length vector using an internal representation. The decoder generates the caption by reading the encoded vectors. Merge model, as described by [26], has been used in the proposed model where the encoded form of an image is merged with the encoded form of text and caption generated so far. The combined encoded input is then used by the decoder to generate the next word in the sequence.

4 Dataset

We are working on Gurmukhi text which is a script used for writing the Punjabi language in Punjab and the Northern part of India. As no standard dataset has been available for the task in hand, the required data has been collected from publically available resources like newspaper websites. Dataset of sports domain has been collected from Punjabi Tribune newspaper[1]. This dataset comprises 10,000 images paired with one caption and associated news document each. Dataset is divided into 8000 samples of Training Data, 1000 samples of Validation Data, and 1000 samples of Testing Data. See table 1 for example.

5 Proposed Model

Our approach is the generalization of the Many to Many sequence generation model. Encoder decoder architecture has been successfully applied to such problems [23]. The aim here is to generate caption from image looking cues from the associated news text. Thus we need to integrate image features with text features. In literature, a number of methods are mentioned to incorporate image features with text features. Some combine image feature with word feature, some use image feature as attention while other uses image feature as a part of the input sequence. [25] and [24] presents a comprehensive overview on this. For the task in hand we tried various models as follow:

5.1 Model 1(Baseline)

The baseline model is a simple model using VGG-19 network on encoder side and unidirectional LSTM on decoder side (see figure 3). We have used a pre-trained VGG-19 model as a convolution neural network to extract the features of the image. It maps the image to a fixed-length vector. 4096 one-dimensional image feature vector is extracted from the fc7 layer of the VGG-19 network.

$$I = VGG(Images) \tag{1}$$

where I is the features of the image.

Each caption is concatenated with a special word "Startseq" at the beginning and with "Endseq" at the end to make the network recognize the

[1] "https://www.punjabitribuneonline.com/"

Caption: ਭਾਰਤੀ ਹਾਕੀ ਖਿਡਾਰੀ ਮਨਦੀਪ ਸਿੰਘ ਨਿਊਜ਼ੀਲੈਂਡ ਦੇ ਮਾਰਕਸ ਚਾਈਲਡ ਨਾਲ ਭਿੜਦਾ ਹੋਇਆ (Indian hockey player Mandeep Singh clashes with New Zealand's Marcus Child)

News: ਭਾਰਤ ਨੇ ਆਪਣਾ ਸ਼ਾਨਦਾਰ ਪ੍ਰਦਰਸ਼ਨ ਜਾਰੀ ਰੱਖਦਿਆਂ ਨਿਊਜ਼ੀਲੈਂਡ ਨੂੰ ਅੱਜ 4-0 ਗੋਲਾਂ ਨਾਲ ਹਰਾ ਕੇ ਤਿੰਨ ਟੈਸਟ ਮੈਚਾਂ ਦੀ ਲੜੀ 3-0 ਨਾਲ ਜਿੱਤ ਲਈ। ਭਾਰਤ ਨੇ ਪਹਿਲੇ ਮੈਚ ਵਿੱਚ ਨਿਊਜ਼ੀਲੈਂਡ ਨੂੰ 4-2 ਗੋਲਾਂ ਨਾਲ ਅਤੇ ਦੂਜੇ ਮੈਚ ਵਿੱਚ 3-1 ਗੋਲਾਂ ਨਾਲ ਹਰਾਇਆ ਸੀ। ਭਾਰਤ ਦੀ ਜਿੱਤ ਵਿੱਚ ਰੁਪਿੰਦਰਪਾਲ ਸਿੰਘ ਅੱਠਵੇਂ ਮਿੰਟ ਸੁਰਿੰਦਰ ਕੁਮਾਰ 15ਵੇਂ ਮਿੰਟ ਮਨਦੀਪ ਸਿੰਘ 44ਵੇਂ ਮਿੰਟ ਅਤੇ ਆਕਾਸ਼ਦੀਪ ਸਿੰਘ 60ਵੇਂ ਮਿੰਟ ਨੇ ਗੋਲ ਕੀਤੇ। ਭਾਰਤ ਨੇ ਇਸ ਤਰ੍ਹਾਂ ਵਿਸ਼ਵ ਦੀ ਨੌਵੇਂ ਨੰਬਰ ਦੀ ਟੀਮ ਨਿਊਜ਼ੀਲੈਂਡ ਨੂੰ ਤਿੰਨ ਮੈਚਾਂ ਵਿੱਚ ਹਰਾ ਕੇ ਅਗਲੇ ਮਹੀਨੇ ਹੋਣ ਵਾਲੀਆਂ ਏਸ਼ਿਆਈ ਖੇਡਾਂ ਲਈ ਆਪਣੀਆਂ ਮਜ਼ਬੂਤ ਤਿਆਰੀਆਂ ਦੇ ਸੰਕੇਤ ਦਿੱਤੇ ਹਨ । ਰੁਪਿੰਦਰ ਨੇ ਪਹਿਲੇ ਕੁਆਰਟਰ ਵਿੱਚ ਪੈਨਲਟੀ ਕਾਰਨਰ ਨੂੰ ਗੋਲ ਵਿੱਚ ਬਦਲ ਕੇ ਟੀਮ ਨੂੰ ਲੀਡ ਦਿਵਾਈ। ਟੂਰਨਾਮੈਂਟ ਵਿੱਚ ਇਹ ਉਸ ਦਾ ਚੌਥਾ ਗੋਲ ਸੀ। ਰੁਪਿੰਦਰ ਨੇ ਇਸ ਤੋਂ ਬਾਅਦ ਸੁਰਿੰਦਰ ਲਈ ਮੌਕਾ ਬਣਾਇਆ, ਜਿਸ 'ਤੇ ਉਸ ਨੇ ਗੋਲ ਕਰ ਦਿੱਤਾ। ਮਿਡਫੀਲਡਰ ਵਿੱਚ ਅਨੁਭਵੀ ਸਰਦਾਰ ਸਿੰਘ ਅਤੇ ਸਿਮਰਜੀਤ ਸਿੰਘ ਨੇ ਇਸ ਤੋਂ ਬਾਅਦ ਖੱਬੇ ਪਾਸਿਓਂ ਮਨਦੀਪ ਸਿੰਘ ਲਈ ਮੌਕਾ ਬਣਾਇਆ । ਜਿਸ 'ਤੇ ਉਨ੍ਹਾਂ ਨੇ ਨਿਊਜ਼ੀਲੈਂਡ ਦੇ ਗੋਲਕੀਪਰ ਜੋਰਜ ਇਨਰਸੈਨ ਨੂੰ ਚਕਮਾ ਦਿੰਦਿਆਂ ਗੋਲ ਕਰਕੇ ਟੀਮ ਦੀ ਲੀਡ ਨੂੰ 3-0 ਗੋਲ ਕਰ ਦਿੱਤਾ। ਸਟਰਾਈਕਰ ਆਕਾਸ਼ਦੀਪ ਸਿੰਘ ਨੇ ਆਖ਼ਰੀ ਹੂਟਰ ਵੱਜਣ ਤੋਂ ਠੀਕ ਪਹਿਲਾਂ ਇੱਕ ਹੋਰ ਗੋਲ ਕਰਕੇ ਨਿਊਜ਼ੀਲੈਂਡ ਦੇ ਜ਼ਖ਼ਮਾਂ 'ਤੇ ਲੂਣ ਛਿੜਕਣ ਦਾ ਕੰਮ ਕੀਤਾ । ਭਾਰਤੀ ਟੀਮ ਦੇ ਮੁੱਖ ਕੋਚ ਹਰਿੰਦਰ ਸਿੰਘ ਨੇ ਇੰਡੋਨੇਸ਼ੀਆ ਵਿੱਚ 18 ਅਗਸਤ ਤੋਂ ਦੋ ਸਤੰਬਰ ਤੱਕ ਹੋਣ ਵਾਲੀਆਂ ਏਸ਼ਿਆਈ ਖੇਡਾਂ ਤੋਂ ਪਹਿਲਾਂ ਟੀਮ ਦੇ ਪ੍ਰਦਰਸ਼ਨ 'ਤੇ ਤਸੱਲੀ ਪ੍ਰਗਟਾਈ। ਉਨ੍ਹਾਂ ਕਿਹਾ ''ਦੁਨੀਆਂ ਦੀਆਂ ਚੋਟੀ ਦੀਆਂ 10 ਟੀਮਾਂ ਵਿੱਚ ਸ਼ੁਮਾਰ ਨਿਊਜ਼ੀਲੈਂਡ ਤੋਂ ਇਹ ਲੜੀ ਜਿੱਤਣ ਨਾਲ ਏਸ਼ਿਆਈ ਖੇਡਾਂ ਲਈ ਸਾਡੀਆਂ ਤਿਆਰੀਆਂ ਨੂੰ ਮਜ਼ਬੂਤੀ ਮਿਲੇਗੀ। ਇਨ੍ਹਾਂ ਤਿੰਨਾਂ ਮੈਚਾਂ ਦੌਰਾਨ ਅਸੀਂ ਵੱਖ ਵੱਖ ਢੰਗ ਅਪਣਾਏ ਅਤੇ ਪੈਨਲਟੀ ਕਾਰਨਰ 'ਤੇ ਵੀ ਵੱਖ ਵੱਖ ਖਿਡਾਰੀਆਂ ਨੂੰ ਪਰਖਿਆ। ਅਸੀਂ ਹੁਣ ਅਗਲੇ ਟੂਰਨਾਮੈਂਟ ਲਈ ਪੂਰੀ ਤਰ੍ਹਾਂ ਤਿਆਰ ਹਾਂ ।'' ਕੋਚ ਨੇ ਨਾਲ ਹੀ ਕਿਹਾ ''ਅਸੀਂ ਮੈਦਾਨੀ ਗੋਲ ਕਰਨ ਵਿੱਚ ਅਜੇ ਹੋਰ ਸੁਧਾਰ ਕਰ ਸਕਦੇ ਹਾਂ ਅਸੀਂ ਅੱਜ ਕੁੱਝ ਆਸਾਨ ਮੌਕੇ ਗੁਆਏ ਅਤੇ ਏਸ਼ਿਆਈ ਖੇਡਾਂ ਤੋਂ ਪਹਿਲਾਂ ਅਸੀਂ ਇਸ ਪਹਿਲੂ 'ਤੇ ਕੰਮ ਕਰਾਂਗੇ ''।

Translation:(Continuing their impressive run, India beat New Zealand 4-0 today to win the three-Test series 3-0. India had beaten New Zealand 4-2 in the first match and 3-1 in the second match. Rupinderpal Singh scored in the 8th minute, Surinder Kumar in the 15th minute, Mandeep Singh in the 44th minute and Akashdeep Singh in the 60th minute. India have signaled their strong readiness for next month's Asian Games by beating world number nine New Zealand in three matches. Rupinder converted a penalty corner in the first quarter to give the team the lead. This was his fourth goal of the tournament. Rupinder then created an opportunity for Surinder, on which he scored. Veteran midfielders Sardar Singh and Simarjit Singh then created an opportunity for Mandeep Singh from the left. He dodged New Zealand goalkeeper George Inersen to give the team a 3-0 lead. Striker Akashdeep Singh added another goal just before the last hoot to sprinkle salt on New Zealand's wounds. Indian head coach Harinder Singh expressed satisfaction over the team's performance ahead of the Asian Games to be held in Indonesia from August 18 to September 2. "Winning the series from New Zealand, one of the top 10 teams in the world, will strengthen our preparations for the Asian Games," he said. During these three matches, we used different methods and also tested different players on penalty corners. We are now fully prepared for the next tournament. "The coach added:" We can make further improvements in field goal scoring.)

Table 1: Sample image, news and caption from dataset

Figure 2: Feature extraction model of Images

starting and end of the caption. LSTM model is used on the decoder side. The model is provided with image features and "Startseq" token to generates the next word. This word is combined with the previous caption and passed to the model as input to generate the next word. Thus it is a recursive process and continues till the token "Endseq". Word embeddings of size 200 were used. Each embedding is further passed to the LSTM layer with 256 hidden units that extract the features of each word. Dropout of 0.5 has been applied to the LSTM layer. The training was done using the teacher enforcement technique. The beam search method with beam size 3 has been used to select the final output.

5.2 Model 2 (Multi-model Architecture)

This model extended the baseline system by utilizing the features of text with images. The news article provides the context information or background knowledge of the image. Thus features of text were extracted using a bi-directional LSTM network. We propose to directly maximize the

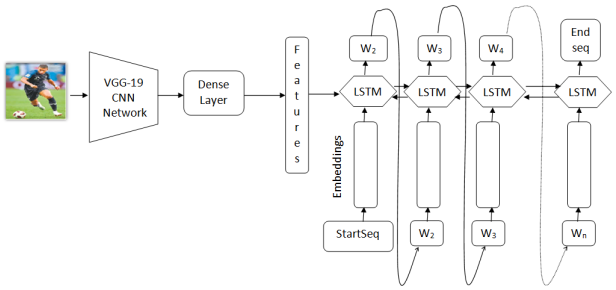


Figure 3: Architecture for Caption Generation from Image

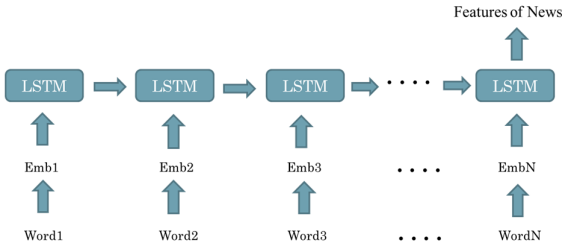


Figure 4: Feature extraction model of Text Document

probability of the caption given the image and its associated text. It is given as,

$$\theta^* = argmax\left(\sum_{(I,D,C)}(\log p(C|I,D;\theta)\right) \quad (2)$$

Where θ is the parameter of the model, I denotes an image, D denotes the document, C is the caption. Let the length of the caption be N, so to model the joint probability over $C_0, ..., C_N$:

$$\log p(C|I,D) = \sum_{t=0}^{N}(\log\ p(C_t|I,D,C_0,...,C_{t-1}) \quad (3)$$

Where C_t is the word at time step t and I is the features of the image. Word embeddings are created for each word in the text document, which converts each word into fixed-size vectors of size 200 units. Each embedding is further passed to the Bi-LSTM layer that learns the features of the news text at each time step. Each LSTM cell has 128 hidden units and a dropout of 0.5 has been used after the Bi-LSTM layer. The features of the image and text are combined and used to initialize the decoder. The decoder is an LSTM network that is provided with one word "Startseq". The next word is generated by combining information from the previous caption. Thus it is a recursive process and continues till the token "Endseq" appears.

5.3 Model 3 (Multi-model architecture with Attention)

Attention is proved to have a positive impact in sequence to sequence task. The encoder-decoder model is extended with Luong's dot attention [17] to get the weightage of source words on target selection. The output of LSTM is combined with context which was calculated using Luong's dot product method. The combined context is passed through a dense layer producing the target word. These features are utilized to calculate context vector as per equation

$$Score(h_t, \overline{h_s}, f_i) = h_t^T.\overline{h_s}.f_i \quad (4)$$

$$\alpha_{ts} = \frac{exp\ (Score(h_t, \overline{h_s}, f_i)}{\sum_{s'=1}^{S} exp\ (Score(h_t, \overline{h_s}, f_i)} \quad (5)$$

$$c_t = \sum_s \alpha_{ts}.\overline{h_s} \quad (6)$$

Where f_i is image features, α_{ts} is attention vector and c_t is context vector. This context vector is concatenated with the output of decoder LSTM to predict the next token in sequence (see figure 5). All other setup remain same.

5.4 Model 4 (Bridging Source and Target)

Following [13], we can shorten the distance between the source and target words and thus strengthen the association, by bridging source and target word embeddings. This model tries to bridge the target side with the source side by determining the most likely source word aligned to it and use the word embedding of this source word to support the prediction of the target hidden state of the next target word to be generated. The rest of the settings are the same as the previous model.

5.5 Model 5 (Keyword Augmentation)

It was observed that due to the sparseness of the data model is unable to produce specific names during caption generation. Inducing some keywords from external sources may improve the output of the model. Keywords can be extracted from the associated News articles and provided to the model instead of starting from token "Startseq". Off the

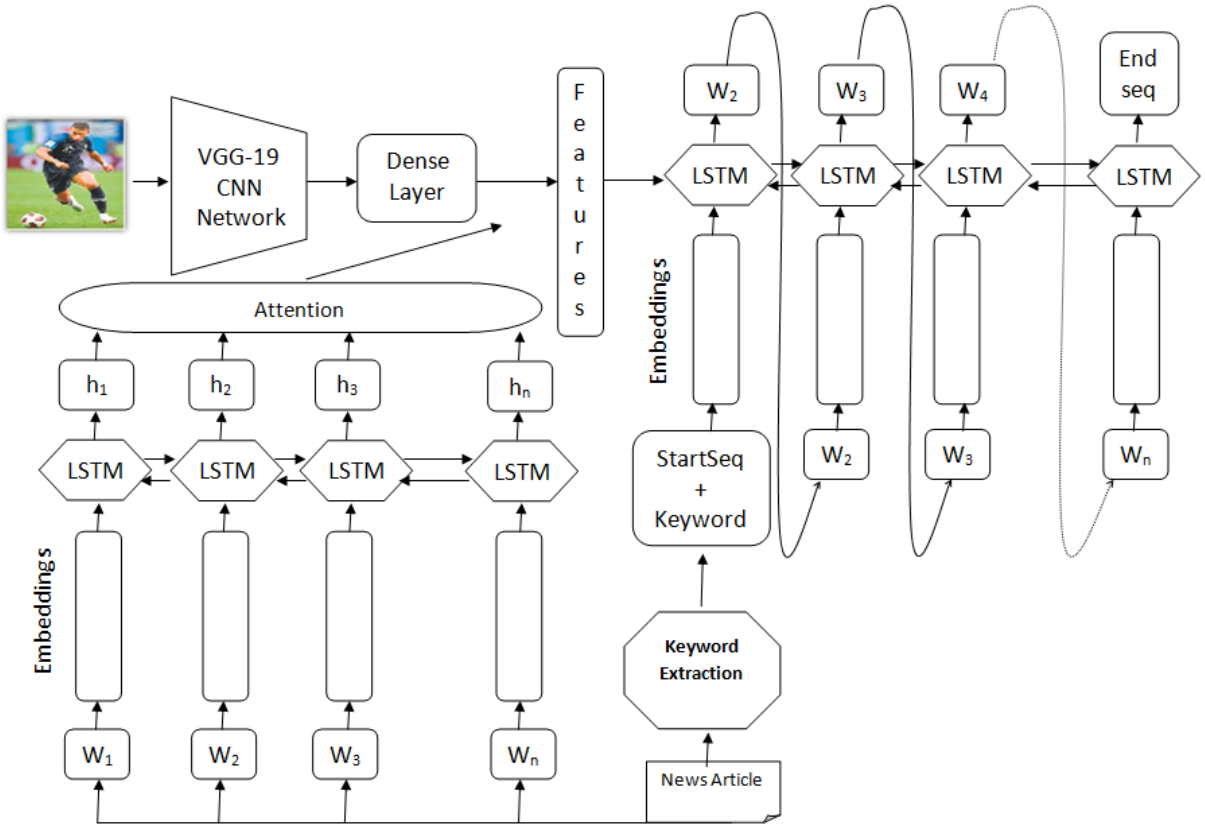


Figure 5: Proposed Architecture for Caption Generation

shelf keyword extractor "RAKE" has been used for extracting keywords. RAKE is short for Rapid Automatic Keyword Extraction algorithm as described by [21]. It needs a list of stop words. The list provided by [20] has been used. The top five keywords were selected and three outputs for each keyword were generated using model 3. These outputs were again ranked based on cosine similarity (see equ. 7) between caption and news text and one top caption is selected as final output.

$$\cos(\mathbf{d}, \mathbf{c}) = \frac{\mathbf{dc}}{\|\mathbf{d}\|\|\mathbf{c}\|} = \frac{\sum_{i=1}^{n} \mathbf{d}_i \mathbf{c}_i}{\sqrt{\sum_{i=1}^{n} (\mathbf{d}_i)^2} \sqrt{\sum_{i=1}^{n} (\mathbf{c}_i)^2}} \quad (7)$$

Where d is news article and c is caption.

6 Experimental Setup

6.1 Preprocessing

It optimizes the performance of the network. Captions and News text are cleaned to remove the data which is not required for training such as punctuation mark, numbers, any meaningless characters, etc. As CNN accepts the fixed-size input, so images are reshaped to the size 224 X 224. Vocabulary for this system is created by considering only unique words from the training captions and news document. Vocabulary thus formed consists of 55943 words.

6.2 Evaluation Metrics

Evaluation of our system is done on the basis of widely used parameters i.e. BLEU Score. BLEU (Bilingual Evaluation Understudy) is an evaluation metric that measures the quality of the caption generated by the particular model against the reference caption. BLEU score is always between 0 and 1. BLEU score does not consider grammaticality, suitability, and intelligibility of the text. BLEU score works best when multiple reference captions are available. As we have only one caption against each test image, the BLEU score may not reveal the performance of the system accurately. Thus human

Table 2: Interpretation table for Fleiss' kappa

Range of K	Interpretation
0.01 – 0.20	Poor agreement
0.21 – 0.40	Slight agreement
0.41 – 0.60	Fair agreement
0.61 – 0.80	Moderate agreement
0.81 – 1.00	Almost perfect agreement

Table 3: BLEU score of different models

Model	BLEU Score
Model 1	26.71
Model 2	27.25
Model 3	28.43
Model 4	28.52
Model 5	29.44

evaluation has been employed to check the suitability of captions generated by the system. Human evaluators judge the quality of the generated captions by the system against the given image and news document by rating on a 1-5 rating scale (1- poor, 2- fair, 3- good, 4- very good, 5- best). Five human evaluators were asked to rate the captions for intelligibility and suitability. Intelligibility tells how intelligible the caption is and Suitability defines how suitable is caption with the given image and news document. Human evaluators checked 100 images selected randomly from test data. All evaluators were presented with same set of 100 images. The inter-rater consistency of human evaluators is checked using Fleiss' kappa. It calculates the degree of agreement in rating over the one that would be expected by chance. For interpreting kappa 'k' values, Table 1 has been used.

7 Results

Table 3 compares the BLEU Scores of captions generated by different models. It can be observed that the proposed system has been reported the highest BLEU score over the scores of others. Model 3 produces a better score than model 2. This confirms that augmenting text data with an image helps in producing better captions. Model 4 performs almost the same as model 3 with a BLUE score marginally more than model 3. Thus bridging the source and target side does not contribute much in this case. Finally, keyword induction yields the highest BLEU score of 29.44. This confirms the fact that the associated news features can improve the caption of the image. An image can have a number of captions but each test image has only one caption in reference, so the BLEU score is low. A Low BLEU score implies that the generated caption didn't match with the reference caption. But this does not mean that caption is wrong. So we perform a human evaluation to judge the quality and suitability of the generated caption. Nevertheless, the BLEU score clearly shows the improvement of the proposed model over the base model.

Table 4: Human evaluation of different models

Model	Intelligibility	Suitability
Model 1	44.65%	19.25%
Model 2	67.88%	28.43%
Model 3	72.00%	48.74%
Model 4	70.58%	47.71%
Model 5	88.21%	58.74%

Table 4 clearly shows that the performance of the proposed system improves by providing keywords as seed words for the caption generation. The Intelligibility of the baseline system is 44.65% which improves to 88.21% in the proposed model. This is a well-known fact that the LSTM network is capable of producing an intelligible sentence from a small set of training corpus. Although captions generated by the baseline model are intelligible they are more general in nature. So we ask human evaluators to judge the captions for suitability also.

Table 4 also shows that although captions generated by the baseline system are intelligible they are poorly suited to the images. The suitability of the baseline system is 19.25% whereas the proposed system reports the caption as 58.74% suitable. This clearly shows that associated text influences the generation of the caption. We use the limited vocabulary for training and the system fails to identify new words in the text. More improvement can be expected by incorporating word embeddings of the whole vocabulary. Many other factors affect the performance of the proposed model as a small dataset which results in small vocabulary and only one caption per image hinders the system to learn from the image in different ways. Fleiss' kappa values are shown in Table 5. Based on kappa values shown in table 5, we find that kappa value in all the cases i.e., Intelligibility and Suitability lies in the range of 0.81-1.00, which clearly shows that evaluators are having the same consent regarding ratings. According to the Table 5, almost perfect agreement is between all the evaluators, hence, justifying our evaluated results

Table 5: Calculation results of Fleiss' kappa

Model	Intelligibility	Suitability
Model 1	0.82343	0.82343
Model 2	0.83257	0.83257
Model 3	0.83316	0.94326
Model 4	0.82343	0.96830
Model 5	0.83257	0.93522

8 Error Analysis

Table 6 shows the output of various models for a given image and news article. All the outputs are quite intelligible as shown in table 6. But the suitability of caption generated by the baseline model is low. The model is not able to extract the context of the image and produces the wrong caption. The image is depicting a football match whereas the caption is about the cricket match. Augmenting image with text produce better caption but it is general in nature. Attention mechanism further improves the system by selecting more suitable words from the article. The caption generated by the proposed model is most suitable as per the human evaluators.

9 Conclusion

In this paper, the deep neural network has been analyzed to generate the captions. This system is trained on images, its captions, and its news document. We propose to augment the system with keywords from within the news article to improve the suitability of the caption. Our proposed model is able to achieve a BLEU score of 29.44. The human evaluation shows that the proposed model is capable of generating more intelligible and suitable caption. It is shown by our results that having hardware limitation and less dataset, vocabulary, it is still possible to learn and generate an abstractive caption. Although we experimented on the sports domain and in the Punjabi language, the proposed model can be trained for any kind of domain and can be extended for any language. As a part of future work, we can increase the vocabulary by collecting more datasets. Moreover, the vocabulary of this system depends on the words available in the system dataset only, universal vocabulary can be used for better results. Pre-trained word embeddings can enhance the performance of the system. Improving the data set in terms of a number of associated captions for training will certainly put a positive impact on performance.

References

[1] Asawa, J., Deshpande, M., Gaikwad, S., and Toshniwal, R. Caption recommendation system. *United International Journal for Research & Technology (UIJRT)*, 2:4–9, 2021.

[2] Batra, V., He, Y., and Vogiatzis, G. Neural caption generation for news images. In chair), N. C. C., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Hasida, K., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., Piperidis, S., and Tokunaga, T., editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 7-12, 2018 2018. European Language Resources Association (ELRA).

[3] Biten, A. F., Gomez, L., Rusinol, M., and Karatzas, D. Good news, everyone! context driven entity-aware captioning for news images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[4] Chen, J., Dong, W., and Li, M. Image caption generator based on deep neural networks, 2016.

[5] Devlin, J., Gupta, S., Girshick, R. B., Mitchell, M., and Zitnick, C. L. Exploring nearest neighbor approaches for image captioning. *CoRR*, abs/1505.04467, 2015.

[6] Elamri, C. and Planque, T. Automated neural image caption generator for visually impaired people, 2016.

[7] Faruk, A. M., Faraby, H. A., Azad, M. M., Fedous, M. R., and Morol, M. K. Image to bengali caption generation using deep cnn and bidirectional gated recurrent unit. In *2020 23rd International Conference on Computer and Information Technology (ICCIT)*, pages 1–6, 2020.

[8] Feng, Y. and Lapata, M. How many words is a picture worth? automatic caption generation for news images. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1239–1249, Uppsala, Sweden, July 2010. Association for Computational Linguistics.

	News: ਫਰਾਂਸ ਨੇ ਰੂਸ ਵਿੱਚ 21ਵਾਂ ਫੀਫਾ ਵਿਸ਼ਵ ਕੱਪ ਜਿੱਤ ਕੇ ਪੰਜ ਮੁਲਕਾਂ ਅਰਜਨਟੀਨਾ ਜਰਮਨੀ ਬ੍ਰਾਜ਼ੀਲ ਇਟਲੀ ਤੇ ਯੁਰੂਗੁਏ ਦੀ ਬਰਾਬਰੀ ਕਰ ਲਈ ਹੈ ਜਿਨ੍ਹਾਂ ਨੇ ਘੱਟੋ ਘੱਟ ਦੋ ਦੋ ਵਾਰ ਵਿਸ਼ਵ ਖ਼ਿਤਾਬ ਜਿੱਤੇ ਹਨ ਇਸ ਟੂਰਨਾਮੈਂਟ ਦੌਰਾਨ ਫਰਾਂਸ ਨੇ ਦੱਖਣੀ ਅਮਰੀਕਨ ਟੀਮਾਂ ਬ੍ਰਾਜ਼ੀਲ ਤੇ ਅਰਜਨਟੀਨਾ ਅਤੇ ਯੂਰੋਪੀਅਨ ਟੀਮਾਂ ਜਰਮਨੀ ਤੇ ਇਟਲੀ ਦੀ ਸਰਦਾਰੀ ਨੂੰ ਚੁਣੌਤੀ ਦਿੱਤੀ ਹੈ ਪਿਛਲੇ 30 ਵਰ੍ਹਿਆਂ ਦੌਰਾਨ ਫਰਾਂਸ ਨੇ ਫੁਟਬਾਲ ਦੀਆਂ ਸਿਖਰਾਂ ਨੂੰ ਛੂਹਿਆ ਹੈ ਇਸ ਦੌਰਾਨ ਫਰਾਂਸ ਇੱਕ ਵਾਰ ਉਪ ਜੇਤੂ ਅਤੇ ਇੱਕ ਵਾਰ ਤੀਜੇ ਅਤੇ ਚੌਥੇ ਸਥਾਨ 'ਤੇ ਰਹਿ ਚੁੱਕਿਆ ਹੈ ਫਰਾਂਸ ਨੇ ਪਿਛਲੇ ਤਿੰਨ ਦਹਾਕਿਆਂ ਦੌਰਾਨ ਦੋ ਵਾਰ ਯੂਰੋ ਕੱਪ ਵੀ ਜਿੱਤਿਆ ਜੋ ਦੂਜਾ ਸਭ ਤੋਂ ਵੱਡਾ ਫੁਟਬਾਲ ਟੂਰਨਾਮੈਂਟ ਹੈ ਇਸ ਟੂਰਨਾਮੈਂਟ ਵਿੱਚ ਉਹ ਇੱਕ ਵਾਰ ਉਪ ਜੇਤੂ ਤੇ ਇਕ ਵਾਰ ਸੈਮੀ ਫਾਈਨਲ ਤੱਕ ਪਹੁੰਚਿਆ ਹੈ ਫਰਾਂਸ ਨੇ ਵਿਸ਼ਵ ਫੁਟਬਾਲ ਨੂੰ ਜ਼ਿਨੇਦਿਨ ਜ਼ਿਦਾਨ ਪਲਾਤਿਨੀ ਡਿਡਿਅਰ ਡੀਸ਼ਾਂ ਥਇਏਰੀ ਹੈਨਰੀ ਫੈਬੀਅਨ ਬਾਰਥੇ ਜਿਹੇ ਖਿਡਾਰੀ ਦਿੱਤੇ ਮੌਜੂਦਾ ਟੀਮ ਦੇ ਐਂਟਨੀ ਗ੍ਰੀਜ਼ਮੈਨ ਤੇ ਕਿਲੀਅਨ ਮਬਾਪੇ ਵਿਸ਼ਵ ਫੁਟਬਾਲ ਦੇ ਚਮਕਦੇ ਸਿਤਾਰੇ ਹਨ ਬ੍ਰਾਜ਼ੀਲ ਦੇ ਪੇਲੇ ਤੋਂ ਬਾਅਦ ਮਬਾਪੇ ਫੁਟਬਾਲ ਜਗਤ ਦਾ ਦੂਜਾ ਛੋਟੀ ਉਮਰ 19 ਸਾਲ ਦਾ ਖਿਡਾਰੀ ਹੈ ਜਿਸ ਨੇ ਵਿਸ ਕੱਪ ਦੇ ਫਾਈਨਲ ਵਿੱਚ ਗੋਲ ਕੀਤਾ ਟੂਰਨਾਮੈਂਟ ਦੌਰਾਨ ਚਾਰ ਗੋਲ ਕਰਨ ਵਾਲਾ ਮਬਾਪੇ ਫਰਾਂਸ ਦੀ ਭਵਿੱਖ ਦੀ ਸਭ ਤੋਂ ਵੱਡੀ ਉਮੀਦ ਅਤੇ ਵਿਸ਼ਵ ਫੁਟਬਾਲ ਦਾ ਨਵਾਂ ਸਿਤਾਰਾ ਬਣ ਗਿਆ ਹੈ **Translation:**France have won the 21st FIFA World Cup in Russia, leveling with five nations, Argentina, Germany, Brazil, Italy and Uruguay, who have won the World Cup at least twice. During the tournament, France has challenged the dominance of South American teams Brazil and Argentina and European teams Germany and Italy. France has reached the pinnacle of football in the last 30 years. France, meanwhile, have been runners-up once and third and fourth once. France has also won the Euro Cup twice in the last three decades, the second largest football tournament. He has been a one-time runner-up and a semi-finalist in the tournament. France gave world football players such as Zinedine Zidane, Platini, Didier Dixon, Thierry Henry, Fabian Barthe. The current team's Anthony Griezmann and Killian Mbabane are the shining stars of world football. After Pel of Brazil, Mbabane is the second youngest player in football to score in a World Cup final. Mbabane, who scored four goals during the tournament, has become France's biggest hope for the future and a new star in world football.
Model 1	Caption: ਏਬੀ ਡਿਵੀਲੀਅਰਜ਼ ਆਖਰੀ ਦਿਨ ਦੌੜਾਂ ਬਣਾਉਂਦਾ ਹੋਇਆ (AB deVilliers making runs on last day)
Model 2	Caption: ਮੈਚ ਦੌਰਾਨ ਭਿੜਦੇ ਹੋਏ ਖਿਡਾਰੀ (Players clashing during match)
Model 3	Caption: ਮੈਚ ਦੌਰਾਨ ਸ਼ਾਟ ਲਾਉਂਦਾ ਹੋਇਆ ਫਰਾਂਸ ਦਾ ਕੁਮਾਰ ਸੰਗਾਕਾਰਾ (France's Kumar Sangakara playing a shot during match)
Model 4	Caption: ਦੱਖਣੀ ਅਮਰੀਕਨ ਦਾ ਖਿਡਾਰੀ ਮੈਚ ਦੌਰਾਨ ਸ਼ਾਟ ਲਾਉਂਦਾ ਹੋਇਆ (South Africa's batsman playing a shot during match)
Model 5	Caption: ਫੀਫਾ ਮੈਚ ਦੌਰਾਨ ਸ਼ਾਟ ਜੜਦਾ ਹੋਇਆ Keyword: ਫੀਫਾ (Playing a shot during FIFA match)

Table 6: Sample image, news and captions generated by different models

[9] Feng, Y. and Lapata, M. Automatic caption generation for news images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(4):797–812, April 2013.

[10] Hossain, M. Z., Sohel, F., Shiratuddin, M. F., and Laga, H. A comprehensive survey of deep learning for image captioning. *ACM Comput. Surv.*, 51(6), Feb. 2019.

[11] Karpathy, A. and Li, F. Deep visual-semantic alignments for generating image descriptions. *CoRR*, abs/1412.2306, 2014.

[12] Kiruthika, N., Devi, V., and selvi M.N.D, T. Extractive and abstractive caption generation model for news images. *International Journal of Innovative Research in Technology & Science(IJIRTS)*, 2(2):80–86, 2012.

[13] Kuang, S., Li, J., Branco, A., Luo, W., and Xiong, D. Attention focusing for neural machine translation by bridging source and target embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1767–1776, Melbourne, Australia, jul 2018. Association for Computational Linguistics.

[14] Kulkarni, G., Premraj, V., Ordonez, V., Dhar, S., Li, S., Choi, Y., Berg, A. C., and Berg, T. L. Babytalk: Understanding and generating simple image descriptions. *IEEE Transac-*

tions on Pattern Analysis and Machine Intelligence, 35(12):2891–2903, Dec 2013.

[15] Kuznetsova, P., Ordonez, V., Berg, A. C., Berg, T. L., and Choi, Y. Collective generation of natural image descriptions. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers - Volume 1*, ACL '12, page 359–368, USA, 2012. Association for Computational Linguistics.

[16] Li, S., Tao, Z., Li, K., and Fu, Y. Visual to text: Survey of image and video captioning. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 3(4):297–312, Aug 2019.

[17] Luong, M., Pham, H., and Manning, C. D. Effective approaches to attention-based neural machine translation. *CoRR*, abs/1508.04025, 2015.

[18] Mason, R. and Charniak, E. Domain-specific image captioning. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*, pages 11–20, Ann Arbor, Michigan, June 2014. Association for Computational Linguistics.

[19] Ordonez, V., Kulkarni, G., and Berg, T. L. Im2text: Describing images using 1 million captioned photographs. In Shawe-Taylor, J., Zemel, R. S., Bartlett, P. L., Pereira, F., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems 24*, pages 1143–1151. Curran Associates, Inc., 2011.

[20] Puri, R., Bedi, R. S., and Goyal, D. V. Automated stopwords identification inpunjabi documents. *Research Cell: An International Journal of Engineering Sciences*, 8(1):119–125, 2013.

[21] Rose, S., Engel, D., Cramer, N., and Cowley, W. *Automatic Keyword Extraction from Individual Documents*, pages 1 – 20. Wiley, 03 2010.

[22] Solari, F., Wang, H., Zhang, Y., and Yu, X. An overview of image caption generation methods. *Computational Intelligence and Neuroscience*, 2020:1–13, 2020.

[23] Sutskever, I., Vinyals, O., and Le, Q. V. Sequence to sequence learning with neural networks. *CoRR*, abs/1409.3215, 2014.

[24] Tanti, M., Gatt, A., and Camilleri, K. P. What is the role of recurrent neural networks (rnns) in an image caption generator? *CoRR*, abs/1708.02043, 2017.

[25] Tanti, M., Gatt, A., and Camilleri, K. P. Where to put the image in an image caption generator. *CoRR*, abs/1703.09137, 2017.

[26] TANTI, M., GATT, A., and CAMILLERI, K. P. Where to put the image in an image caption generator. *Natural Language Engineering*, 24(3):467–489, 2018.

[27] Vinyals, O., Toshev, A., Bengio, S., and Erhan, D. Show and tell: A neural image caption generator. *CoRR*, abs/1411.4555, 2014.

[28] Vinyals, O., Toshev, A., Bengio, S., and Erhan, D. Show and tell: A neural image caption generator. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3156–3164, June 2015.

[29] Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., and Bengio, Y. Show, attend and tell: Neural image caption generation with visual attention. In Bach, F. and Blei, D., editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 2048–2057, Lille, France, 07–09 Jul 2015. PMLR.

[30] Yang, Z. and Okazaki, N. Image caption generation for news articles. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1941–1951, Barcelona, Spain (Online), Dec. 2020. International Committee on Computational Linguistics.

[31] Yao, B. Z., Yang, X., Lin, L., Lee, M. W., and Zhu, S. I2t: Image parsing to text description. *Proceedings of the IEEE*, 98(8):1485–1508, Aug 2010.

[32] You, Q., Jin, H., Wang, Z., Fang, C., and Luo, J. Image captioning with semantic attention. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4651–4659, June 2016.

[33] Zhang, S., Zhang, Y., Chen, Z., and Li, Z. Vsam-based visual keyword generation for image caption. *IEEE Access*, 9:27638–27649, 2021.

[34] Zhao, S., Sharma, P., Levinboim, T., and Soricut, R. Informative image captioning with external sources of information. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6485–6494, Florence, Italy, July 2019. Association for Computational Linguistics.