

Deep learning approach for Image captioning in Hindi language

MSc Research Project
Data Analytics

Ankit Rathi
Student ID: x18104207

School of Computing
National College of Ireland

Supervisor: Mr. Manuel Tova-Izquierdo



National College of Ireland
Project Submission Sheet
School of Computing

National
College of
Ireland

Student Name:	Ankit Rathi
Student ID:	x18104207
Programme:	Data Analytics
Year:	2019
Module:	MSc Research Project
Supervisor:	Mr. Manuel Tova-Izquierdo
Submission Due Date:	12/08/2019
Project Title:	Deep learning approach for Image captioning in Hindi language
Word Count:	8519
Page Count:	25

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	
Date:	August 10, 2019

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	<input type="checkbox"/>
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	<input type="checkbox"/>
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	<input type="checkbox"/>

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Deep learning approach for Image captioning in Hindi language

Ankit Rathi
x18104207

Abstract

Generating image description automatically from the content of an image is one of the fundamental problem in artificial intelligence. This task involves the knowledge of both computer vision and natural language processing, called “Image caption generation”. Many research has been carried out in this field, but it was mainly focused on generating image descriptions in English, as existing image caption datasets are mostly in English. However, the image description generator model should not be limited by language. The lack of image captioning dataset other than English is a problem, especially for a morphologically rich language such as Hindi. Thus, this research constructed Hindi image description dataset based on images from Flickr8k dataset using Google cloud translator, which is called “Flickr8k-Hindi Datasets”. The Flickr8k-Hindi Datasets consist of four datasets based on a number of description per image and clean or unclean descriptions. The study uses these Hindi datasets to train encoder-decoder neural network model. The experiments showed that training the model with a single clean description per image generates high-quality caption than a model trained with five uncleaned descriptions per image. Although model trained with five uncleaned descriptions per image achieved BLEU-1 score of 0.585, which is the current state of art.

Keywords— Image captioning, Hindi language, Monolingual learning, encoder-decoder framework, Neural network.

1 Introduction

In recent years, automated description of the image has received a lot of attention. Many researchers worked in image captioning field and showed a significant improvement, especially in the encoder-decoder framework of image captioning (Vinyals et al.; 2015). The encoder-decoder framework is the most popular and effective method to generate image description, which involves neural networks such as Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN). In this process, CNN and RNN-LSTM (Long Short Term Memory) are used as an encoder to encode images and text data respectively, and in the decoder, language neural network model is used to merge both data and predict the sequence of words (Vinyals et al.; 2015) (Karpathy and Fei-Fei; 2015) (Johnson et al.; 2016) (You et al.; 2016). There are number of datasets available for English captioning e.g., Flickr8K (Hodosh et al.; 2013), Flickr30K (Young et al.; 2014), MSCOCO (Lin et al.; 2014) and so on. Because of lack of datasets in another language, there are few works in image captioning other than English. Some researchers have collected image

description data other than English, for example, Japanese, Chinese, French, German, Dutch and Spanish. The main objective of their work was to collect image description from crowdsourcing and secondary objective was to build an monolingual image captioning model (Yoshikawa et al.; 2017) (Li et al.; 2016) (Elliott et al.; 2016) (van Miltenburg et al.; 2017). The possible application of image captioning in the native language can be useful for image search, description of the image in multilingual e-commerce website, social media, news website, and audio-described image for visually impaired people.

Image captioning has been done in three most popular languages in the world that are Chinese, Spanish and English (Gomez-Garay et al.; 2018). But there is no research of image captioning in Hindi language, which is the fourth most spoken language of the world (Verma and Singh; 2018). Also, it is the most widely spoken language in the Indian subcontinent, and some people living in that region don't understand English (Nair et al.; 2016). Hence image captioning in Hindi can be very useful to the people who don't know other languages except Hindi. Thus, it is very necessary to explore: ***How image captioning model works for Hindi description dataset?***

There are number of datasets available for English image captioning, among them, Flickr8k dataset (Hodosh et al.; 2013) is selected as a base dataset of the research because it is the smallest available dataset, which includes 8000 images and 40,000 descriptions. The main object of the research is to generate an image description in Hindi. To accomplish the objective, it is essential to have image description data in Hindi. Therefore, this project constructed a Hindi version of Flickr8K dataset by translating English description of Flickr8k dataset to Hindi using Google Cloud Translator API ¹. The new dataset is named as “Flickr8k Hindi Datasets”². The unique thing about this dataset is that it consists of four datasets based on a number of description and cleanliness of the data i.e. unclean-5 sentences, unclean-1 sentence, clean-5 sentences, clean-1 sentence. From the name “unclean-5 sentences” dataset, it consists of five uncleaned descriptions per image. Similarly, “clean-1 sentence” dataset consists of single cleaned description per image and so on. The encoder-decoder neural network model is trained with all four datasets and quality of generated caption is measured by human evaluation and machine evaluation using BLEU (Bilingual Evaluation Understudy) (Papineni et al.; 2002).

In previous work of image captioning other than English language, most of the researchers have collected the image description from crowd workers ((Miyazaki and Shimizu; 2016), (Li et al.; 2016)). Furthermore, image captioning model trained with description collected from crowdsourcing generated more natural caption (Yoshikawa et al.; 2017). But due to the limitation of time, this project is not able to collect image description from Hindi speaking crowd, and alternatively, machine translator is used. To the best of my knowledge, this is the first paper to generate image description in Hindi. Furthermore, the image captioning model of this research is well suited as a baseline for future work on image caption generation in Hindi.

The paper is organised as follow: Section 2 explains past work on image captioning in English and Non-English language. Then in section 3, the methodology of the project is explained with design process flow and model architecture. After that section 4 presents how the dataset is created and how model is trained to generate image caption in Hindi. Next, the result of the experimental evaluation is presented in section 5. Finally, section 6 and section 7 includes the discussion and conclusion part of the project.

¹<https://cloud.google.com/translate/>

²<https://github.com/rathiankit03/ImageCaptionHindi/tree/master/Flickr8kHindiDataset>

2 Related Work

This section discusses the state of art and the different strategies applied by the authors in the field of image captioning. Recently, many authors have worked in this field and proposed different methodologies to describe the content of the picture in English or non-English language. Depending on the language of an image description, the section is further divided into subsections. In each subsection, the various methodologies or techniques applied by authors are analysed and implemented approach is justified with supporting evidence.

2.1 Image captioning in the English Language

Since 2010, many research groups worked on image captioning and reported a significant improvement (Bai and An; 2018). Most of the research used English language dataset to train their image captioning model. This development was made possible due to most of the available dataset for image captioning was in English. In these 9 years, several technique or methodologies have been implemented to describe an image in natural language. Initial work of image captioning was mainly focused on two lines of research method, i.e. retrieval based, and template-based (Bai and An; 2018). These methods were dependent on hard-coded language structure, the disadvantage was that they were not flexible enough (Bai and An; 2018). As a result, these methods become extinct and deep neural network (DNN) came into play (Hossain et al.; 2018). In computer vision (C.V.) and natural language processing (NLP) field powerful DNN has shown significant results. Particularly in image captioning problem, DNN has demonstrated a state of art result (Bai and An; 2018). With the recent surge of research interest in image captioning using DNN, many different approaches are used to describe an image. This section analyses the past work based on DNN approach for image captioning problem, i.e. multimodal learning, Encoder-decoder framework and attention guided framework. The section is further divided into subsection according to the neural network approach of image captioning and last part of the section justifies the use of an appropriate method implemented in this approach.

2.1.1 Multimodal Learning

The initial work using DNN was proposed by Kiros et al. (2014a). The method uses multimodal log-bilinear model, in which CNN is used first to extract the features of images and this feature is forwarded to the neural language model, which uses multimodal space to map image features with text feature and predict the word conditioned on image feature and previously generated word. Considering the limitation of Kiros et al. (2014a) model to handle a large amount of data, Mao et al. (2014) used the Recurrent Neural Network(RNN) as a neural language model. The neural language model works inefficiently to work with long term memory (Hossain et al.; 2018).

2.1.2 Encoder-Decoder framework

To overcome the limitation of the neural language model, Kiros et al. (2014a) extended their work in Kiros et al. (2014b) where Long Short Term Memory(LSTM) is used for sentence encoding. This was the first approach to use the popular encoder-decoder framework in image captioning research and it showed significant results. Although the encoder-

decoder framework of image captioning is inspired from the neural machine translation model proposed by Cho et al. (2014), where the encoder-decoder framework is used to translate the text from one language to another language. In image captioning model of Kiros et al. (2014b), to encode image features and textual data CNN and LSTM is used respectively, and to decode visual elements conditioned on text feature vector neural language model is used. With the same inspiration, Vinyals et al. (2015) used CNN as an encoder to encode images and RNN-LSTM to decode image features into text. In Vinyals et al. (2015)'s encoder-decoder framework, image captioning is formulated as predicting the probability of sentence conditioned on input image feature. Similar to Vinyals et al. (2015) work, Donahue et al. (2014) adopted the same encoder-decoder framework. Instead of inputting image features to the system at the initial stage, Donahue et al. (2014) provide both image and text feature to the sequential language model at each time step. Aiming to improve the image captioning model of the encoder-decoder framework, Jia et al. (2015) extracted the semantic features from images and added the features to each unit of RNN-LSTM during the process of caption generation.

2.1.3 Attention Mechanism

An attention guided framework is the next advance version of the encoder-decoder framework. The first attention mechanism in image caption generator is proposed by Xu et al. (2015). In the attention mechanism, the encoder-decoder framework is used, and while generating an image description, the model is more focused on the salient region of an image. It is a method which has the ability to weight the region of an image differently. For example, it can add more weights to the important region of an image. Xu et al. (2015)'s attention model was adding weights to a random part of an image. Therefore, some important part of an image was getting missed to generate caption. To overcome this limitation, You et al. (2016) introduced a semantic attention model which focuses on linguistically important objects or action in the image. In the above attention mechanism, model force visual attention to be active for each word, even for words which does not explain visual information. For instance stop words like 'the', 'of' etc does not explain the image object. To overcome this problem, Lu et al. (2017) came with a more advanced version of an attention mechanism called an adaptive attention mechanism, which automatically decides whether to rely on the visual signal or language model. If adaptive attention model chooses to attend visual signal, then the model automatically decides which part of the image to attend. Chen et al. (2017) used semantic attention mechanism and compared pre-trained CNN, such as VGG-16 and ResNet50. There experiments shows that using ResNet-50 as CNN gives better result compare to VGG-16. In recent year, Anderson et al. (2017) introduced a combined bottom-up and top-down visual attention mechanism. The bottom-up mechanism group of the salient image region is represented by pooled convolutional feature vector. On other side top-down method uses task-specific content to predict attention distribution over the image regions. This advanced method of Anderson et al. (2017) generate very natural captions and achieved the highest accuracy in image captioning of MS COCO dataset with BLEU-4 score of 0.369.

Thus, there are three main approaches to caption an image i.e. Multimodal method, Encoder-Decoder method and Attention mechanism. To find the best approach of image captioning, evaluation results of all research for Flickr 8K dataset is compared in the

image captioning survey of Bai and An (2018). From this survey, an attention mechanism is an efficient method to caption an image. Basically, the attention mechanism is an encoder-decoder framework where decoder focuses on a specific feature of an image and text at each time step (Jia et al.; 2015). Although attention mechanism is the most dominating method but adding attention to encoder-decoder model increases the complexity and require high computational power (Bai and An; 2018). An Encoder-Decoder framework gives a notable result with less complexity (Li et al.; 2019). This research aims to caption an image in the Hindi language. In past, no research has been done in this topic and results of the encoder-decoder framework to generate a caption in Hindi can be helpful in image captioning domain. Thus, this work considers to implement an encoder-decoder model where pretrained CNN is used to encode the image features and LSTM-RNN is used to encode text feature.

2.2 Image Captioning other than English language

Most of the work on image captioning has considered English as the target language, although some researches have used Japanese, Chinese, German, French, Dutch, Spanish and Czechia as target language. This section analyses the existing theories proposed by various authors and justify the use of the appropriate technique that will be implemented in the project. This section is again broken into different subsection based on the language used in the research.

2.2.1 Image description in Japanese

Most of the available dataset of image captioning was in English. Considering this problem Miyazaki and Shimizu (2016) constructed “YJ caption 26k Dataset”, the first non-English caption dataset. YJ caption 26k is a Japanese version of MS COCO dataset in which captions are collected from crowdsourcing. To find the best way to describe an image in Japanese language, Miyazaki and Shimizu (2016) compared three learning techniques i.e. monolingual, alternative and transfer learning. From there research, they found transfer learning technique is the best method to generate a caption in the Japanese language. Similarly, Yoshikawa et al. (2017) developed “STAIR Caption”, Japanese caption dataset based on images from MS-COCO dataset. Like “YJ caption 26k” dataset, image description in “STAIR caption” dataset is collected from crowdsourcing. In “STAIR caption”, 820,310 Japanese captions is provided for all images of MS-COCO (164,062 images). However, “YJ caption” consist of 131,740 Japanese captions for only 26,000 images of MS-COCO dataset. Thus, “STAIR caption” is the largest Japanese image caption dataset. Yoshikawa et al. (2017) compared their model trained with “STAIR caption” dataset and machine translated dataset and in result, they found model trained with “STAIR caption” dataset generate more fluent captions. Tsutsui and Crandall (2017) used “YJ caption 26K” dataset for multilingual image captioning. They trained their single model with two languages (dual-language model) i.e. English and Japanese, and inserted artificial tokens at the beginning of the sentence to switch the language of the caption. Along with the dual-language model, Tsutsui and Crandall (2017) also trained their model with Japanese caption only (monolingual model) and after evaluating the caption generation they found monolingual model performs better than dual-language model.

From the above research, it is clear that caption dataset is required to develop image

captioning model other than the English language. There are two ways to develop a caption dataset; a) Collecting captions from crowdsourcing ((Miyazaki and Shimizu; 2016), (Yoshikawa et al.; 2017)), b) Collecting captions using machine translation (Yoshikawa et al.; 2017). In Japanese captioning, a model trained with crowdsource caption were able to generate more natural captions with good accuracy (Yoshikawa et al.; 2017). The above works are thoroughly analysed, and in result it has been found that model performance is more dependent on description data collected from crowd workers. For instance, Miyazaki and Shimizu (2016) and Yoshikawa et al. (2017) used monolingual model in there experiment and both achieved different results, because both authors used different dataset collected from various group of people. Hence, the quality of caption collected from crowdsourcing may vary and it can affect the accuracy of the model (Yoshikawa et al.; 2017). The same thing has been observed in Chinese caption generation which is given in the next section. Overall, in all three research of Japanese caption generation monolingual model has given excellent performance. Therefore, this research proposed to implement a monolingual model trained with Hindi caption.

2.2.2 Image description in Chinese

In Japanese captioning research, image descriptions were collected from two sources i.e., crowdsource and machine translator. However to generate Chinese caption, Li et al. (2016) collected image description of Flicker 8k images from three sources i.e., crowdsource, machine translator and human translator. Caption collected from crowdsourcing is named as “Flicker 8k-CN”. While collecting the caption from crowdsourcing Li et al. (2016) noticed the cultural gap in the collected description. For instance, the English caption describes a picture of the woman as an Asian woman, but Chinese caption collected from Chinese speaking population describes the same picture of the woman as a middle-aged lady. From this observation, Li et al. (2016) conclude that there can be a cultural difference when captions are collected from crowdsourcing located in a different geographical region. They also compared the accuracy of a model trained with the machine-translated caption, crowdsource caption and human translated caption. In result Li et al. (2016) found the model trained with machine-translated captions outperforms the model trained with the human translated caption, and model trained with crowdsource caption is at the top in terms of accuracy. The reason behind this result was that crowdsource captions are more fluent and natural compared to machine translated caption (Li et al.; 2016). Therefore, Lan et al. (2017) proposed “Fluency guided framework” where they developed fluent machine-translated image description by editing non-fluent machine translated sentence manually. In their approach, the model trained with “Fluency guided framework” outperformed the model trained with machine-translated sentence. To support this line of research in Chinese captioning Li et al. (2019) developed “COCO-CN” dataset, the Chinese version of MSCOCO dataset collected from crowdsourcing.

Similar to Japanese work ((Miyazaki and Shimizu; 2016), (Yoshikawa et al.; 2017)), Chinese authors ((Li et al.; 2016) and (Li et al.; 2019)) also developed caption dataset from crowdsourcing. Although the model trained with crowdsource caption has given more natural caption, but it requires a large amount of time to collect. Apart from crowdsource caption, Li et al. (2016) and Lan et al. (2017) used the human resource to translate and edit the image description respectively. But model trained with human translated sentences has not given considerable result (Li et al.; 2016) and to develop the dataset it also requires a large amount of time and money. In Japanese captioning Yoshikawa

et al. (2017) observed different captioning result collected from different crowdsource. The similar observation has been seen by Chinese author (Li et al.; 2016), and they conclude that cultural gap is the reason behind the different captioning result. Thus, to sum up, caption collected from crowdsourcing can improve the image captioning, but it also takes a large amount of time to collect and it is also limited to the region from where crowdsourcing caption is collected. Considering this as big disadvantage, the other simple solution is to collect the image description by translating the English caption dataset to the target language. Li et al. (2016), Li et al. (2017) and Lan et al. (2017) used an image captioning model trained with machine-translated caption and from there experiment they found model can generate considerable captions. Thus, this project uses a machine-translated caption to generate image description in Hindi.

2.2.3 Image Description in European languages

Other than Japanese and Chinese captioning, few research has been done to generate image description in German, French, Dutch and Spanish language. Elliott et al. (2015) proposed a multilingual image captioning model, where English and German text are grounded parallelly against image feature. To train the multilingual model, they used images aligned with the German caption of IAPR-TC12 dataset (Grubinger et al.; 2007). To allow further research in German image captioning, Elliott et al. (2016) developed “Multi30k” dataset for Flickr30k image data by collecting German image descriptions from crowdsourcing and professional human translators. In their research, they noticed that translated sentence collected from human translators have the same number of tokens compared to English caption dataset, and image description collected from crowd workers were different in terms of length and word type. To extend the Multi30K dataset Elliott et al. (2016) also added a French description. There are two evaluations method to evaluate image captioning model i.e. machine evaluation method and human evaluation method. Elliott et al. (2016) compared this evaluation method by evaluating multimodal translation and multilingual image captioning model. The main finding of their research is: human evaluation evaluates the generated caption more naturally and accurate compared to machine evaluation metrics.

van Miltenburg et al. (2017) contribute their work in image captioning field to generate image description in the Dutch language. They collected Dutch image description from crowdsourcing and merged with Multi30k Dataset. van Miltenburg et al. (2017) compared the Dutch image description with English and French description in Multi30k, and they found that the description of images was different because of cultural difference. Similar observation has been found in Chinese (Li et al.; 2016) and Japanese captioning (Yoshikawa et al.; 2017). Recently, van Miltenburg et al. (2017) released Dutch image description data, named as DIDECC (The Dutch Image Description and Eye-tracking Corpus) to caption an image in Dutch language. On other side, for visually impaired people, Gomez-Garay et al. (2018) introduced a system to generate and verbalize image descriptions in Spanish.

From the above analysis of research, authors were more focused on collection of the image description. Meanwhile, Elliott et al. (2016) said that human evaluation is the best evaluation method in the field of image captioning. Therefore, the research considers to evaluate generated captions manually. But due to limitation of time and budget, research considers high BLEU scored caption only for human evaluation.

2.3 Conclusion

The past methods and techniques implemented in image captioning research are thoroughly reviewed in this section. Several approaches are used to describe the content of an image in English or non-English language but there is no research carried out to describe an image in Hindi sentence. Therefore, image captioning in Hindi is considered a huge research gap. After analysing the past techniques, the project decides to select the encoder-decoder framework to caption an image in Hindi. The Hindi dataset is required to train the encoder-decoder neural network model but image captioning dataset is not available in the Hindi language. Therefore, considering the time and budget limitation, this study collected the Hindi description by translating the English caption dataset using machine translator. Also in past work, the model trained with machine-translated sentence have shown considerable results. One of the main findings of image captioning research is that evaluating the generated caption by manual observation is the best way of evaluation. Hence, human evaluation is also used in this research. Along with the human evaluation method, BLEU is used as a machine-based evaluation metric.

Thus, to sum up, in all research of non-English image captioning applied monolingual learning model. In the case of Japanese and Chinese captioning, Yoshikawa et al. (2017) and Li et al. (2017) observed different result in their monolingual model. According to there research, more work can be carried out on monolingual image captioning in other language. This previous work concludes that there lies a scope to perform several experiments to generate image description in the Hindi language. The results of this experiment can contribute to the literature on Computer Vision and Natural Language Processing field.

3 Methodology

In data mining related research, usually two methodologies are used to implement the model i.e. CRISP-DM (Cross-Industry Standard Process for Data Mining) (Wirth; 2000) and KDD (Knowledge Discovery and Data Mining) (Mariscal et al.; 2010). The KDD fits as best methodology in proposed approach as the deployment of models in the business layer is not applicable. Therefore, in this project KDD approach is followed to implement the image captioning model. Basically, KDD consist of five stages namely selection, pre-processing, transformation, data mining and evaluation/interpretation (Mariscal et al.; 2010). But this project uses modified KDD shown in the Figure 1. In the image caption generator model, two types of data is required to caption an image i.e. Image data and Text data. Thus, to process and transform both type of data, the modified KDD approach consist of two layers: Visual layer and Language layer. After transformation of visual feature and text feature into a numerical vector, both data is fed to the image captioning model, and at last stage, models are evaluated and interpreted by using evaluation metrics or observing the output.

The past research have convincingly shown that the combination of CNN-RNN can produce a rich description of images. Hence, proposed work has used CNN-RNN model in the encoder-decoder model. Vinyals et al. (2015) finds in there experiment that pretrained CNN improves the performance of the model. Inspiring from there research, this project uses three different pre-trained CNN (VGG-16, ResNet-50 and Inception-V3) and finds the best CNN for Hindi caption generation. In case of RNN, Long Short Term Memory (LSTM) (Hochreiter and Schmidhuber; 1997) is used as a recurrent unit, which has shown

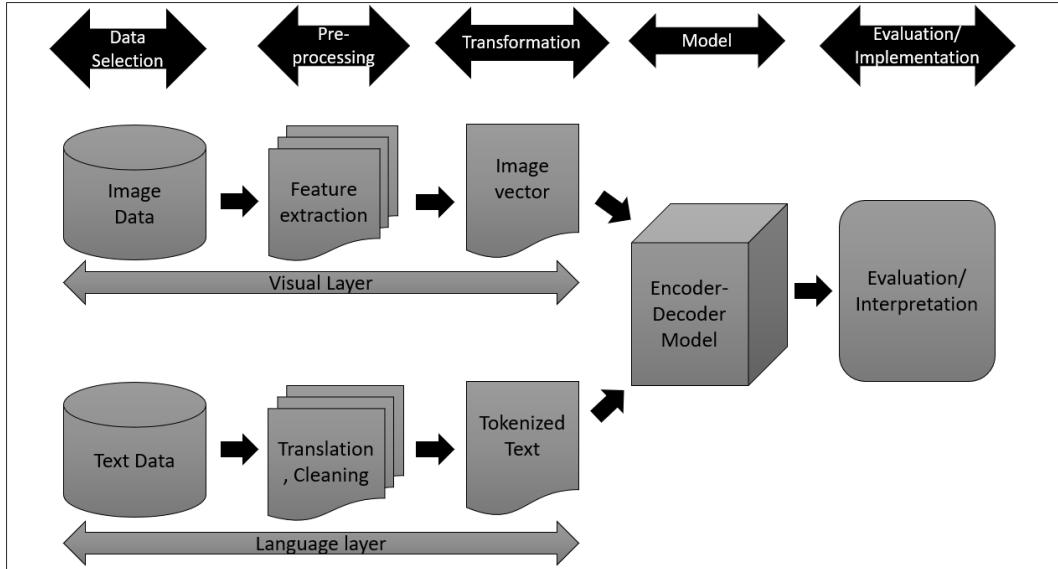


Figure 1: Modified KDD methodology for this research project.

great success in image Captioning field (Bai and An; 2018). Hence, the proposed work uses three neural network model, CNN and LSTM as an encoder to encode the image feature and text feature, and neural language model as a decoder to decode the image feature to the natural sentence. The section is further divided to explain the design process flow and model structure of image captioning model used in this project.

3.1 Design Process flow

The project design process of image captioning is illustrated in below Figure 2. It gives the visual information about how the caption generation process worked in this project. The training descriptions are wrapped between startseq and endseq tags, so that model predicts the caption starting with startseq and end with endseq tag. Following the approach of Vinyals et al. (2015), the model predicts a word conditioned on previous words sequence and image vector. In other words, the model generates a new image description word by word based on given caption prefix and image feature. Both, the caption vector and image vector is passed to the model and it predicts the word which has the highest probability. The predicted word is appended to the caption and passed the same word to the model again. As shown in the Figure 2, it is an iterative process, which will divide into two ways. The first way to stop the caption generation is an artificial tag “endseq”. When the model predicts the “endseq” tag, it stops predicting next word as it is the last word appeared in all training caption. The second way to stop caption generation is by giving an upper bound of caption length as the maximum length of generated description. Using any of the two ways, model terminates and output the generated caption.

3.2 Model Structure

To develop the encoder-decoder neural network model, it is important to select the most suitable model structure (Tanti et al.; 2018). However, there are many research about how to select a good model structure for a given task. Among those research, Tanti et al.

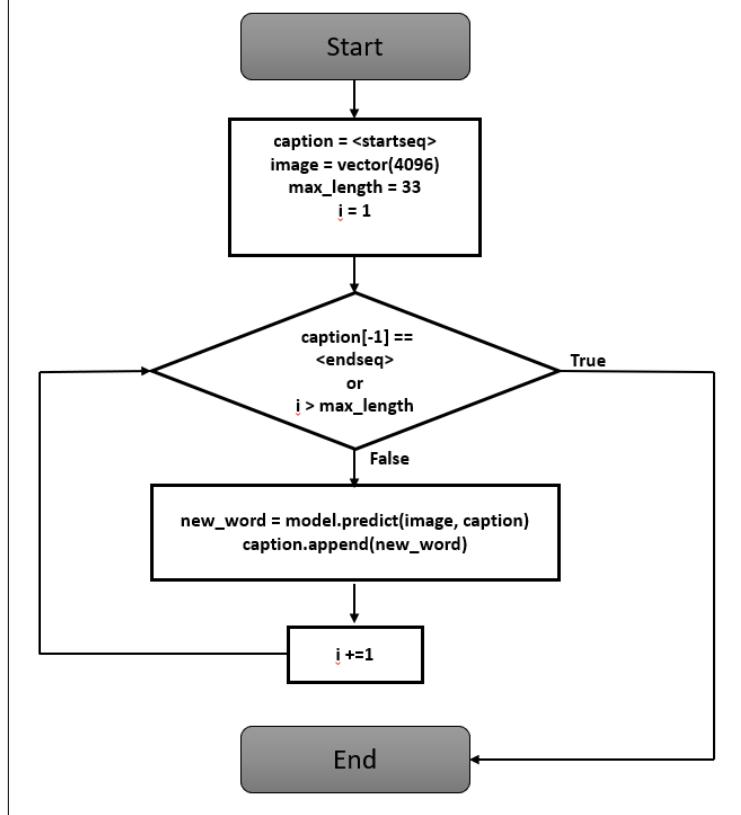


Figure 2: Design process flow of image captioning model used in this project.

(2018) compared 16 different model architecture for image captioning and determined the best model structure i.e. Merge Model. Therefore, Merge model is used in this project and it depicted in the below Figure 3.

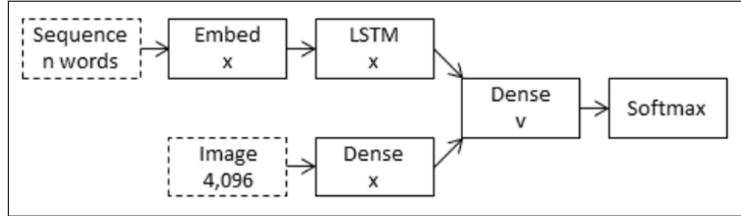


Figure 3: Model structure of encoder-decoder model used in this project.

The merge model takes two vectors as input i.e. Word sequence vector and Image feature vector. The combination of these two vector inputs is then used by the decoder model to generate the next probable word in the sequence. According to Tanti et al. (2018) the model where the text data and image data are handled exclusively perform better. In image caption generation the RNN-LSTM neural network model is the best model to encode the text data (Tanti et al.; 2017). Therefore, the proposed approach has used RNN-LSTM to encode the text data, and pre-trained CNN model is used to encode the image data. Both outputs of encoders are merged and there are multiple ways to combine the two encoded inputs, such as addition, concatenation, and multiplication. Among these methods, experiment by Tanti et al. (2018) has shown addition as the best

method to combine two encoded inputs. Hence, the addition method is used in proposed work to merge the image feature vector and word vector. The combination of both inputs is fed to the dense layer, which is the decoder of the proposed model. In the method of addition, image feature vector and word vector are added together and the resulting vector has the same length to the input vector. The softmax layer is the last layer of the model, and it uses a greedy search to returns a single word which has the highest probability based on the previous word generated and image feature.

3.3 Data Collection

In this research, due to limitation of time and resource, Flickr8k dataset (Hodosh et al. (2013)) is selected as image caption dataset. It is the smallest dataset available for image captioning. Also, it is freely available and can be accessed after submitting the request form to the University of Illinois³. It consists of eight thousand images and five English descriptions per image. The images in dataset are about daily life and the captions of the images are collected from crowdsourcing. The image description of the dataset is mainly describing the key objects and scenes shown in the images.(Hodosh et al., 2013)

4 Implementation

This section focuses on the process of implementing the image captioning model in the Hindi language. It includes the details information with supporting evidence from the starting phase of implementation to the end. The various phases of implementation includes environmental setup, data creation, data pre-processing, data transformation and modelling. All these implementation phases are discussed as follow:

4.1 Environmental Setup

The implementation of the project is performed in a cloud server sponsored by ESDS⁴ , and CentOS Linux-7 is used as an operating system. The data processing and modelling are implemented in Python 3.6. The training of the image captioning model consumes a lot of memory (Vinyals et al.; 2017). Therefore, all the experiments are executed on 64 GB of RAM. The project also used high-level APIs such Keras⁵ and Google Cloud Translation⁶ for modelling and machine translation. The reason behind choosing Keras library is: a) It is easy to understand and fast experimentation, b) Supports both CNN, RNN and combination of two, and c) It runs on CPU (Gulli and Pal; 2017).

4.2 Creation of Hindi dataset

The proposed work aims to caption an image in Hindi, but the available dataset is in English. There are two ways to develop Hindi descriptions; one way is to collect Hindi image description from Hindi speaker crowd and the second way is to translate the caption dataset into Hindi language using machine translator. Considering the time

³<https://forms.illinois.edu/sec/1713398>

⁴<https://www.esds.co.in/>

⁵<https://keras.io/>

⁶<https://cloud.google.com/translate/>

and budget limitation, this project uses machine translator. Google cloud translator⁷ is the best available machine translator, which provide free API to translate 71 languages. In English-Hindi translation Google scored 0.402 ATEC (Assessment of Text Essential Characteristic). It is the highest score achieved in machine translator of English-Hindi translation (Malik and Baghel; 2019). Therefore, this project used the Google cloud translation API to translate the 40,000 sentences into Hindi. To the best of my knowledge, this is the first time to develop Hindi version of Flickr8k caption dataset, which is named as “Flickr8k-Hindi Dataset” caption dataset available at Github⁸. Below Figure 4 shows the sample of the caption dataset describing an image in English and Hindi.

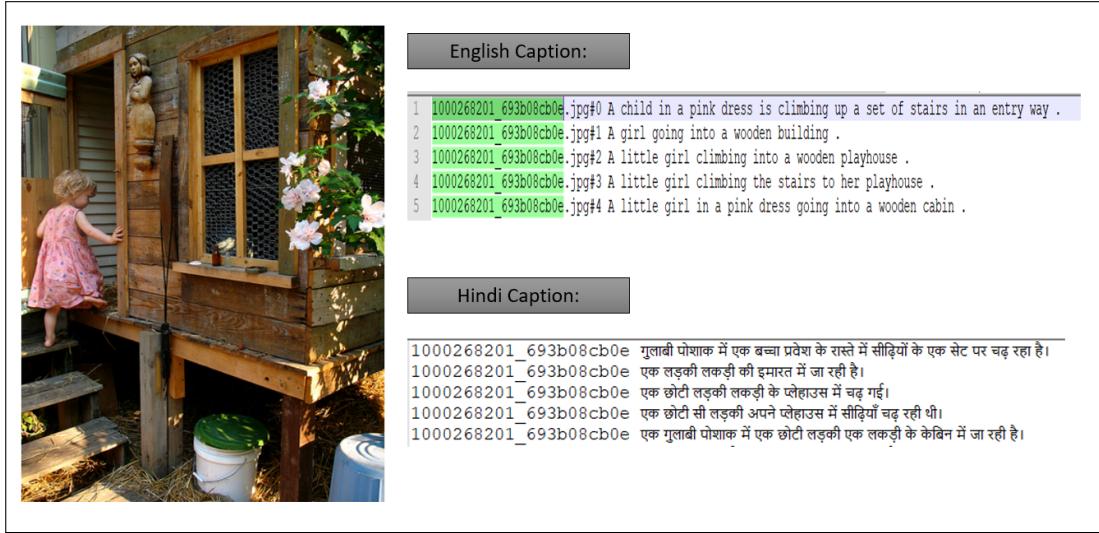


Figure 4: The sample of the caption dataset describing an image in English and Hindi.

To check the performance and quality of the model, four datasets are created from “Flickr8k-Hindi Dataset” i.e Unclean-5 sentences, Unclean-1 sentence, Clean-5 sentences, Clean-1 sentences. From the name of dataset “Unclean-5 Sentence” consists of 5 descriptions per image and non-informative words are not removed from the descriptions. However, in “Clean-1 sentence”, only one description is kept for single image and this description does not contain non-informative words. The same applies for “Unclean-1 sentence” and “Clean-5 sentences”. All the minute details about this datasets are given in Configuration Manual report (Section 3.2).

4.3 Data Processing and Transformation

The Flicker8k-Hindi dataset consists of image data (8,000 images) and text data (40,000 sentences). Both data need to pre-process and transform before fitting into the model. This section explains how image and textual data is prepared before forwarding to the deep neural network image caption generation model.

4.3.1 Image Data

The image needs to be in the form of a vector when it is fed as input to the deep neural network. In the survey of “Deep learning for Image captioning”, Hossain et al. (2018)

⁷<https://cloud.google.com/translate/>

⁸<https://github.com/rathiankit03/ImageCaptionHindi/tree/master/Flickr8kHindiDataset>

observed that to convert images into fixed-sized vector pre-trained CNN has been used in every state of art researches of image captioning. There are many pre-trained CNN model available for image caption generator model, such as AlexNet, VGG-16, GoogleNet, ResNet-50 and Inception-V3 (Bai and An; 2018). Among these pre-trained CNN, VGG-16 (Simonyan and Zisserman; 2014), ResNet-50 (He et al.; 2015) and Inception-V3 (Szegedy et al.; 2016) have shown the significant result for Flickr8k dataset (Bai and An; 2018). Therefore, all these three CNN has been used in this research as a transfer learning model. Also, pre-trained CNN avoids overfitting in image captioning model (Vinyals et al.; 2017). Usually, the last layer of these pre-trained CNN is use to predict the classification of the image. Therefore, the last layer of pre-trained CNN has been removed and the second last layer which returns the feature of image in the form of vector is used while implementing the model. Instead of computing image features at every run of the model, photo features of VGG-16, ResNet-50 and InceptionV3 are saved into the pickle file. This will make the training of image captioning model faster and consume less memory. Thus, three files of image features are created for VGG-16, ResNet-50 and Inception-V3. The details of all three image features are given in Configuration manual report (Section 4).

4.3.2 Text Data

In image captioning model, the text is the only thing which needs to predict. So during the training of the model text will be the target variable that the model learns to predict. Therefore, before giving text input to the neural network, it is important to pre-process the text data and transform it into a numerical form. This section gives more details about the preprocessing and transformation of text data used in the image captioning model.

(a) Data- Pre-processing:

As discussed in the above section, the Flickr8k-Hindi dataset provides the five captions for each image. There are total 10,889 unique words in the Flickr8k-Hindi Dataset, and a hindi word “ek”, which means ‘a’ or ‘one’ is the most common word appeared. This word does not have much information, and removal of this kind of word can reduce the size of the vocabulary, which can be beneficial for model performance. Removal of stop words can decrease the fluency of the generated caption (Lan et al.; 2017). Therefore, stop words are not removed from the text, and all the numeric words and punctuation are removed. After cleaning the description data vocabulary size of word is decreased to 10,870, which is also a big number. So, to have less number of vocabulary size, one caption is taken out of the 5 captions. In result, the unique vocabulary size and a maximum length of four datasets of “Flickr8k-Hindi Dataset” are as follow (Table 1):

Table 1: The vocabulary size and a maximum length of four datasets of “Flickr8k-Hindi Dataset”

Datset	Unique Vocabulary Size	Maximum length
Unclean-5 Sentences (Flickr8k-Hindi dataset).	10,889	39
Unclean-1 Sentence	4,235	34
Clean-5 Sentences	10,870	35
Clean-1 Sentence	4,216	33

(b) Wrapping tags:

After preprocessing, the text data is labelled with the start marker (“startseq”) and end marker (“endseq”) to teach the machine that it is beginning and end of the sentence. The example is shown in below Figure 5. The importance of these tags is already explained in the Project Design flow section 3.1.

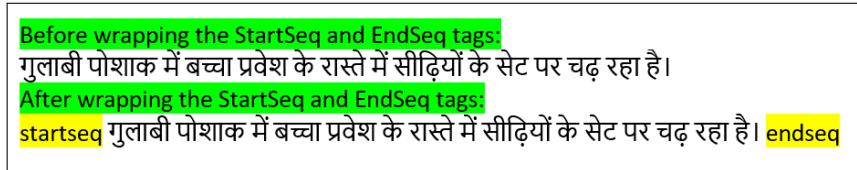


Figure 5: The text data is labelled with the start marker (“startseq”) and end marker (“endseq”) to teach the machine that it is beginning and end of the sentence.

(c) Tokenization:

The neural network cannot work directly on the text. Therefore, after labelling the sentence as mentioned above, each word is converted to integer tokens using Keras tokenizer. But the neural network cannot work on integer either, so the Keras embedding layer is used to convert integer tokens to the vector of floating-point values. The embedding size of this vector is kept similar to the image feature vector by adding zero padding so that they can be merged together and processed as input to the decoder neural network model.

4.4 Experiments Design

The multiple sets of models were trained with different parameters and the best combination of parameter were chosen to train the model on four datasets of “Flickr8k-Hindi Dataset”. The hyperparameter tuning and the training of the model are illustrated as follow:

4.4.1 Tuning of the model

In the “Show and Tell Lesson learned from Image captioning model” paper, Vinyals et al. (2017) took more than one month to optimize the model using GPU(Nvidia-K20). Learning from there experiment, this project used a small sample of the fuller dataset to find the optimal hyperparameter in less time. The small sample of the dataset consists of 200 images with 5 descriptions per image. It is divided equally to training dataset and test dataset. Due to the stochastic nature of the neural network model, variation of the model is studied, and an ideal number of model run repeat is found out by running the model multiple time. From the evaluation of the model, the test harness was stable at 5 model run repeats. Thus, to find the optimal hyperparameter, extensive set of experiment is performed to asses the effectiveness of the model using different model architecture. All the different parameter explored during implementation and their results are given in Google Sheet⁹. After finding the optimal parameter, the captioning model is trained with

⁹https://docs.google.com/spreadsheets/d/1KgNPzXaeCvLN00hwSmjy7PrqK10Peu_u1xDsKCw8rs0/edit?usp=sharing

full dataset. Several CNN's and optimizers were used while performing the experiments. Below are the findings of the experiments:

- (a) Image feature is extracted from the pooling layer and flat layer, in result feature extracted from flat layer benefits the model to predict the caption.
- (b) Extracting image feature from VGG-16 offers an advantage over ResNet-50 and Inception-V3.
- (c) In the case of optimizer, Adam offers benefits over RMS-Prop and Adagrad. Thus, Adam was used as optimizer with default parameter value (learning rate of 0.01 and decay rate of 0.999).
- (d) The models were also compared by adding number of dense layers and LSTM layers. Using multiple layers of dense layer and LSTM layer was giving more error compared to a single layer.
- (e) The different number of neurons in the dense layer and LSTM layer were tested, and in result 128 neurons was find out as an optimal number of neurons.
- (f) The model without the dropout rate was leading to overfit the data. Different combination of dropout rate was tried and in result, 50% of dropout gave improvement in BLEU score.

4.4.2 Training

The image captioning model is trained with four datasets, mentioned in the dataset creation section (Section 4.2). The testing of the optimal hyperparameter (obtained from tuning experiment) is executed by configuring the same parameter with a large dataset. The variation in model performance trained with the full dataset is observed similar to the performance of the model trained with a small dataset. The model configuration experimented in all four datasets is given in Google Sheet¹⁰. It also includes the training time of the model. From the experiment, it has been observed that the model trained with 1-description per image consume less time for training and it also require less memory to execute. During training the model, it is observed that the model was learning fast and overfitting the training set quickly. Therefore, the skill of training model on holdout development dataset was monitored using ModelCheckpoint in Keras, and model is saved when the skill of the model on development dataset improved at the end of an epoch. At the end of run, saved model with the best skill is used as the final model. From the experiments, it has been observed that the model was overfitting below 10 epoch run. Therefore, models were trained for 10 epoch and for every model, loss of training data and developer data is observed by plotting a graph shown in the Figure 6. Below graph (Figure 6) is the example of generated graph while training the final model. From the graph (Figure 6), after 2 epoch, loss of validity dataset is descreases slightly and loss of training data is decreased drastically. It means after 2 epoch, the model is overfit. Also at epoch 6, devolopment data is giving minium error, therefore model is saved at epoch 6 using ModelCheckpoint.

¹⁰<https://docs.google.com/spreadsheets/d/1pXo-IGFeBXFTZLS7P6Ro005pwgLUXCygo17jVqPN5Bg/edit?usp=sharing>

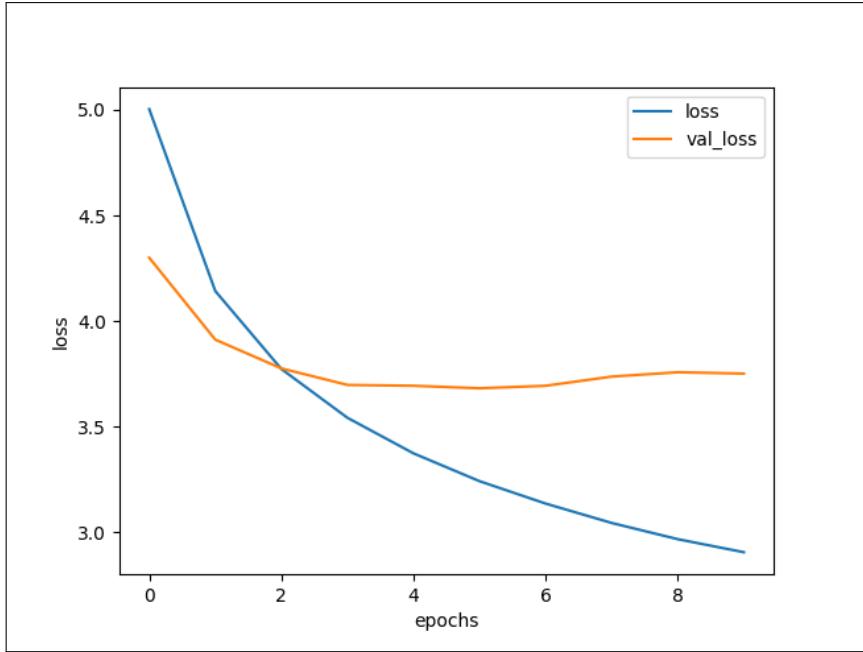


Figure 6: Training of the final model with Clean-1 sentence dataset - after 2 epoch, the model is overfit.

In the experiments, various model is implemented with different model structure and parameter. But this section explains the model structure of final model which has given high-quality image caption. The final image captioning model was trained using mini-batch stochastic gradient descent with fixed learning rate and no momentum, and the gradients were computed by backpropagation. Weights were randomly initialized except for image encoder model because pretrained CNN has been used in encoder model. Structure of fine-tuned image captioning model is explained as follow:

(a) **Image feature input:**

The image encoder model takes an input of 4096-element image feature vector, which was generated using VGG-16. Therefore, the input shape of the image encoder is configured as the same size of the image feature. After the input layer, the dropout layer is added with a dropout rate of 0.5, and the image feature vector is compressed to 128 elements in the dense layer using Rectified Linear Unit (RELU) function.

(b) **Text feature input:**

The text encoder model takes a description of an image as a second input. It expects the input sequence with a maximum length of the description (33 words for Clean-1 sentence dataset), which are passed through encoding layer which uses mask ignore padded values. This is followed by dropout layer with a dropout rate of 0.5. At the last stage, LSTM layer with 128 memory units is used which gives output in 128 element vector.

(c) **Merge and prediction output:**

The decoder model takes input from image encoder model and text encoder model. Both encoders produces 128 element vector, and it is merged using an addition operation. This is then forwarded to a dense layer with 128 neurons. At the final stage, dense layer makes softmax prediction of a word.

5 Evaluation and Results

5.1 Evaluation Metric

5.1.1 BLEU

BLEU (Bilingual Evaluation Understudy) was originally developed to evaluate the machine-translated sentence (Dobre; 2015). It is used to measure the closeness between the reference sentence and the candidate sentence. It rates the quality of the generated description given several reference description by counting n-gram co-occurrences. The longer n-gram scores are used to measure the level of fluency of generated description (Papineni et al.; 2002). The most common practice to use n-gram is up to four (Bai and An; 2018), hence one to four n-gram has been used in this project. BLEU-4 measure the fluency of the generate description while BLEU-1 (unigram) accounts for adequacy. It also measures the precision of the generated description, where precision is the ratio of number of overlapping words to the total words in the candidate sentence (generated description) (Papineni et al.; 2002). BLEU is compatible with multiple languages including Hindi (Joshi et al.; 2013).

5.1.2 Human evaluation

Human evaluation is the best evaluation method to check the quality of the generated image description ((Elliott et al.; 2016), (Vinyals et al.; 2017)). Considering the finding of Vinyals et al. (2017). this project evaluated generated image description by using the image as a reference. Due to the limitation of time, manual evaluation of image descriptions is not possible to rate whole test data (2000 description) for multiple models. Therefore, the generated caption with high BLEU score is evaluated manually to test the quality of the generated caption. It is rated by four scores i.e. Very good, Good, Average, Bad. If the model describes the image without error it is rated as “Very Good”, while “Good” score is assigned to the description which have minor errors. If the description is somewhat related to images, then it is classified as “Average” score. Conversely, if the description is unrelated to the image it is classified as a “Bad” score.

5.2 Results

This section includes the result of the models trained on four datasets (Unclean-5 sentence, Unclean-1 sentence, Clean-5 sentence, Clean-1 sentence). The optimal parameter of the model is obtained from the experiments which were performed for tuning. The section is further divided into two subsections according to the method of evaluation applied and at last, the main findings of the project is explained.

5.2.1 Machine Evaluation results

The Table 2 shows the result of the evaluation metric score of the optimal model trained on 4 datasets. The BLEU-1 score above 0.3 is considered an understandable description, while above 0.5 is considered as fluent description. From the Table 2, we can observe BLEU score of a model trained with 5 descriptions per image is always higher than 1 description per image. The reason is that the candidate sentence has 5 reference sentence to match, therefore BLEU score for model trained with 5 description per image comes high. This project used Vinyals et al. (2017) methodology of the encoder-decoder model. But instead of English caption dataset, this project used Hindi caption dataset. The Table 3 shows the result of Vinyals et al. (2017) model trained on the English version of Flickr8k dataset. Comparing with Vinyals et al. (2017) result, model trained with five Hindi description per image showed significant BLEU score and it is almost similar to the Vinyals et al. (2017)'s result. Thus, this project achieved the state of art result to generate Hindi image description.

Table 2: Results of final model trained on 4 datasets. For unclean-5 sentence dataset, project achieved state of art result.

Dataset	BLEU-1	BLEU-2	BLEU-3	BLEU-4
Unclean-5 Sentences	0.5844	0.4	0.27	0.12
Unclean-1 Sentence	0.19	0.11	0.07	0.028
Clean-5 Sentences	0.4136	0.278	0.201	0.09
Clean-1 sentence	0.304	0.158	0.102	0.04372

Table 3: Result of Vinyals et al. (2017)'s model trained on the English version of Flickr8k dataset.

Author	Dataset	BLEU-1	BLEU-2	BLEU-3
Vinyals et al. (2017)	Flickr8k	0.630	0.424	0.270

5.2.2 Human evaluation results

The Figure 7 shows the result of a human evaluation which was performed by observing the generated caption manually. The generated caption with high BLEU score is collected and each generated caption is rated by observing the images. In the case of models trained with 5-description per image, generated caption with high BLEU-4 score (above 0.3) is evaluated manually. On the other side, if the model is trained with 1-description per image then generated caption with high BLEU-2 score (above 0.3) is evaluated manually. After evaluating the generated caption for four models, it has been observed that the model trained with cleaned dataset generates more natural captions. The caption which is unrelated to the image is rated as "Bad" caption. From the Figure 7, models trained with clean datasets have less ratio of "Bad score" compared to the models trained with unclean datasets. Also, the model trained with "Clean-1 Sentence" dataset have the highest ratio of "Very Good" score, which means it generates high-quality captions. The BLEU score of a model trained with "Clean-1 Sentence" is low, but model generates best quality captions.

The Figure 8 shows the example of rated image description obtained from the model trained with "Clean-1 Sentence" dataset. It is interesting to see, for instance in the first

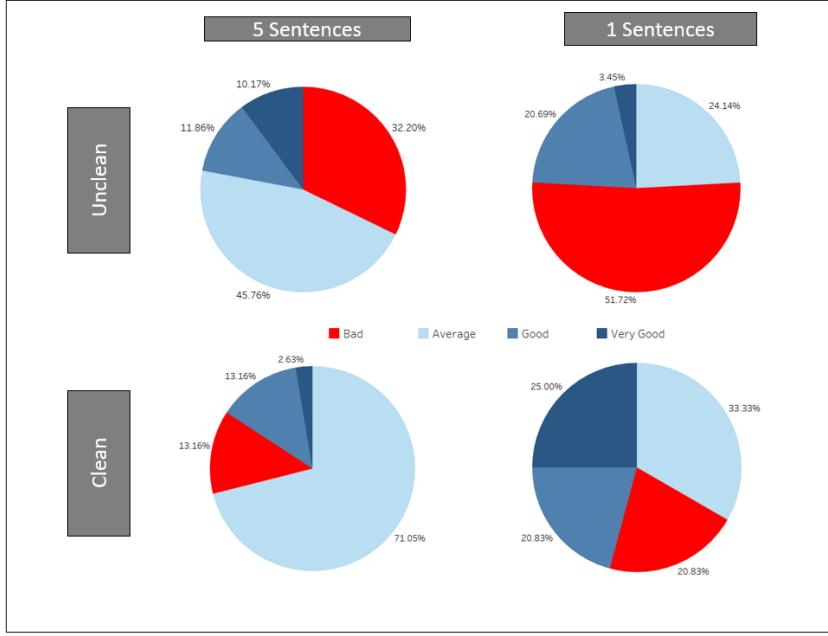


Figure 7: Human evaluation on model trained on 4 datasets. “Very Good” : model describes the image without error; “Good”: model describes the image with minor error; “Average”: description is somewhat related to the image; “Bad”: description is unrelated to the image.

image of the first column, how the model was able to notice a boy jumping into the water. Most of the time model does not predict the exact image description but it describes something related to the image, for example in the first image of the third column, model notice the bike instead of skiing. Some time model predicts the caption which is unrelated to the image. For instance, in the first image of the fourth column, the model describes that man is walking through water which is not related to the image. The working of the model is, it predicts the next probable word conditioned on image feature and previous word. The reason for generating unrelated description may be model is trained with a lot of images which have “man” word in the description. Therefore the probability of “man” word after “startseq” and image feature might be high, and it predicts the “man” instead of “bird”. Thus, model was not able to predict the right caption for that image.

5.2.3 Summary of main findings

The result of the experiments performed during tuning of the model is given in the Implementation section (Section 4.4.1). The main finding of the project is:

- (a) BLEU score of models trained with 5 descriptions per image is higher than models trained with a single description per image (Table 2). But after human evaluation (Figure 7), it has been found that there is no difference between the quality of the generated captions. The reason is that the candidate sentence has 5 reference sentence to match, therefore BLEU score for a model trained with 5 descriptions per image comes high.
- (b) A model trained with cleaned data can generate high-quality caption compare to a

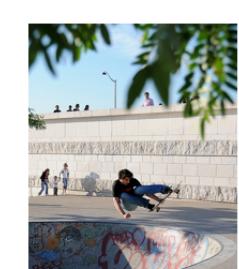
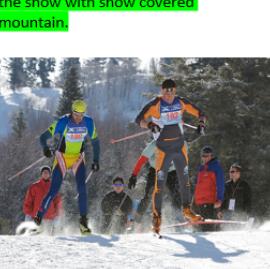
<p>“Very Good” score – Describes without error</p>	 <p>Hindi: लड़का पानी में कूद रहा है। Translation: The boy is jumping into the water.</p>	 <p>Hindi: भेरे रंग का कुत्ता घास में चल रहा है। Translation: The brown dog is walking in the grass.</p>	 <p>Hindi: आदमी अपनो बाइक पर चाल करता है। Translation: The man is doing trick on his bicycle.</p>
<p>“Good” score – Describes with minor error</p>	 <p>Hindi: आदमी बर्फ में बर्फ से ढकी पहाड़ के साथ खड़ा है। Translation: The man stands in the snow with snow covered mountain.</p>	 <p>Hindi: लड़का स्केटबोर्ड से चढ़ता है। Translation: The boy climbs using skateboard.</p>	 <p>Hindi: आदमी पानी के माध्यम से चल रहा है। Translation: Man is walking through the water.</p>
<p>“Average” score – Somewhat related to image.</p>	 <p>Hindi: आदमी बर्फ में बाइक की सवारी कर रहा है। Translation: The man is riding bike in snow.</p>	 <p>Hindi: लाल रंग में आदमी अपने सिर पर फुटबॉल की गेंद को पकड़ने के लिए तैयार हो जाता है। Translation: A man in red gets ready to catch a football ball on his head.</p>	 <p>Hindi: फुटबॉल खिलाड़ी गेंद को पकड़ने के लिए तैयार हो जाता है। Translation: The football player is ready to catch the ball.</p>
<p>“Bad” score – Unrelated to the images</p>	 <p>Hindi: आदमी समुद्र तट पर पानी के माध्यम से चल रहा है। Translation: The man is walking through water on beach.</p>	 <p>Hindi: साइकिल चालक बाइक पर बाइक सवार। Translation: Bike rider on his bike.</p>	 <p>Hindi: आदमी और सफेद कुत्ता पानी में कूद रहा है। Translation: Man and white dog is jumping into water.</p>

Figure 8: Image description obtained from the model trained with “Clean-1 Sentence” dataset.

model trained with unclean data (Figure 7). Because, the frequency of Hindi word “Ek” was very large in unclean data and consequently the probability of Hindi word “Ek” (a/one) next to “startseq” was higher than other words. After removing “Ek” (a/one) from the unclean data, the frequency of the words become balanced in clean dataset. Thus, caption quality is improved in the unclean dataset.

- (c) From the above four models (Figure 7), a model trained with Clean-1 Sentence dataset generates high quality captions and it consumes less memory.

6 Discussion

This study gives an idea about how to generate image captions in Hindi using the encoder-decoder model. After a comparison of four models, the model trained with single cleaned description dataset per image performs significantly and generate high-quality caption. But the BLEU score of that model is low due to less number of reference words to match. This result of the research is very important in the field of image captioning as no study have been done to generate image caption in Hindi. The finding of the study can be useful for further research in Hindi image captioning. As suggested by Vinyals et al. (2017), human evaluation is the best method to evaluate the image caption in English and this seems to be true as evident from the result obtained from this research. Also, the BLEU score of a model trained with Uncleaned-5 sentence dataset came similar to Vinyals et al. (2017)’s model trained with English caption. Thus, this project achieved a state of art result to generate image description in the Hindi language.

The study uses machine-translated sentences to train the image captioning model. It is not necessary that the machine translator returns the translated sentences with one hundred percent accuracy. It faces the challenges to translate the sentence from one language to other languages because of the different grammatical structure of the language. Considering this limitation of machine translator, experiments are performed. Also in previous research, it has been shown that description collected from crowdsourcing generates more natural caption. But due to time limitation, this project was not able to collect the caption from Hindi speaking crowd. The result of this project can be useful for further work who will use crowdsource caption. While evaluating the caption, it has been observed that the same caption was repeating for multiple images. For instance, if the image is about dog doing some activity rather than walking then model generates the caption “Dog is walking through grass”. This type of caption is repeated for multiple images which are related to the dog. The reason is that the model uses a greedy search to predict the next probable word conditioned on image feature and previous word. This method overfits the data quickly and the words which are less appeared in the text have less probability. The solution to this limitation is to use attention mechanism which uses same encoder-decoder model but gives more weights to the relevant image and text. This research was not able to perform experiments on the attention mechanism due to limited source and time. Furthermore, it will be interesting to see how Hindi image captioning will work after applying attention mechanism. Hence in future work, attention mechanism can be applied to generate image description in Hindi.

7 Conclusion

This project has developed image caption dataset for the Hindi language by collecting 40,000 captions for 8,000 images using machine translation. The dataset is named as “Flickr8k-Hindi Dataset”. Under this dataset, four datasets have been created according to number of sentence and cleanliness of the description. This project used these datasets to train encoder-decoder neural network model that can automatically view an image and generate a reasonable description in plain Hindi. The model is trained to predict word by word after receiving the image feature and previous word. Experiments on four datasets showed that the quality of image description is increased after training the model with the clean dataset. Moreover, the model trained with five descriptions per image scored high BLEU score compared to the model trained with one description per image. It is clear from the experiments that the model trained with a single cleaned description per image generates high-quality caption. Also, a model trained with uncleaned five descriptions scored highest BLEU score which is the state of art result to describe an Image in the Hindi language. The results of this experiment can contribute to the literature on Image captioning in Hindi language and it can be used as a baseline model for future research.

8 Acknowledgement

This research was supported by ESDS, cloud data centre in India. I would especially like to thank Mrs Prajakta Jadhav who helped me to get access of cloud server from ESDS. I would also like to extend my thanks to Professor Manuel Tova-Izquierdo for his supervision, guidance and support throughout the research. I would like to acknowledge my father, mother and sister, without their support I couldn't have reached this far.

References

- Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S. and Zhang, L. (2017). Bottom-up and top-down attention for image captioning and visual question answering, *ArXiv170707998 Cs* .
- Bai, S. and An, S. (2018). A survey on automatic image caption generation, *Neurocomputing* **311**: 291–304.
- Chen, Q., Li, W., Lei, Y., Liu, X. and He, Y. (2017). Learning to adapt credible knowledge in cross-lingual sentiment analysis, *Proc. 53rd Annu. Meet. Assoc. Comput. Linguist. 7th Int. Jt. Conf. Nat. Lang. Process* **1**: 419–429.
- Cho, K., Merrienboer, B. V., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H. and Bengio, Y. (2014). Learning phrase representations using rnn encoder-decoder for statistical machine translation., *ArXiv14061078 Cs Stat* .
- Dobre, I. (2015). A comparison between bleu and meteor metrics used for assessing students within an informatics discipline course, *Procedia - Soc. Behav. Sci* **180**: 305–312.

- Donahue, J., Hendricks, L. A., Rohrbach, M., Venugopalan, S., Guadarrama, S., Saenko, K. and Darrell, T. (2014). Long-term recurrent convolutional networks for visual recognition and description, *ArXiv14114389 Cs* .
- Elliott, D., Frank, S. and Hasler, E. (2015). Multilingual image description with neural sequence models, *ArXiv151004709 Cs* .
- Elliott, D., Frank, S., Simaan, K. and Specia, L. (2016). Multi30k: Multilingual english-german image descriptions., *In 5th Workshop on Vision and Language* pp. 70–74.
- Gomez-Garay, A., Raducanu, B. and Salas, J. (2018). Dense captioning of natural scenes in spanish, *Pattern Recognition* pp. 145–154.
- Grubinger, M., Clough, P., Hanbury, A. and Mller, H. (2007). Overview of the imageclefphoto 2007 photographic retrieval task, *presented at the CEUR Workshop Proceedings 1173*.
- Gulli, A. and Pal, S. (2017). *Deep Learning with Keras*, Packt Publishing Ltd.
- He, K., Zhang, X., Ren, S. and Sun, J. (2015). Deep residual learning for image recognition, *ArXiv151203385 Cs* .
- Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory, *Neural Comput.* **9**(8): 1735–1780.
- Hodosh, M., Young, P. and Hockenmaier, J. (2013). Framing image description as a ranking task: Data, models and evaluation metrics, *J. Artif. Intell. Res* pp. 853–899.
- Hossain, M. Z., Sohel, F., Shiratuddin, M. F. and Laga, H. (2018). A comprehensive survey of deep learning for image captioning, *arXiv:1810.04020 [cs, stat]* .
- Jia, X., Gavves, E., Fernando, B. and Tuytelaars, T. (2015). Guiding long-short term memory for image caption generation, *ArXiv150904942 Cs* .
- Johnson, J., Karpathy, A. and Fei-Fei, L. (2016). Densecap: Fully convolutional localization networks for dense captioning, *Presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* pp. 4565–4574.
- Joshi, A., Popat, K., Gautam, S. and Bhattacharyya, P. (2013). Making headlines in hindi: Automatic english to hindi news headline translation, *The Companion Volume of the Proceedings of IJCNLP 2013: System Demonstrations* pp. 21–24.
- Karpathy, A. and Fei-Fei, L. (2015). Deep visual-semantic alignments for generating image descriptions, *Presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* pp. 3128–3137.
- Kiros, R., Salakhutdinov, R. and Zemel, R. (2014a). Multimodal neural language models, *in International Conference on Machine Learning* pp. 595–603.
- Kiros, R., Salakhutdinov, R. and Zemel, R. S. (2014b). Unifying visual-semantic embeddings with multimodal neural language models., *ArXiv14112539 Cs* .

- Lan, W., Li, X. and Dong, J. (2017). Fluency-guided cross-lingual image captioning, *in Proceedings of the 25th ACM International Conference on Multimedia, New York, NY, USA* pp. 1549–1557.
- Li, X., Lan, W., Dong, J. and Liu, H. (2016). Adding chinese captions to images., *In Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval*. pp. 271–275.
- Li, X., Lan, W., Dong, J. and Liu, H. (2017). Add english to image chinese captioning, *IEEE 2nd International Conference on Cloud Computing and Big Data Analysis (ICCCBDA)* pp. 333–338.
- Li, X., Xu, C., Wang, X., Lan, W., Jia, Z., Yang, G. and Xu, J. (2019). Coco-cn for cross-lingual image tagging, captioning and retrieval, *IEEE Trans. Multimed.* .
- Lin, T.-Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., Perona, P., Ramanan, D., Zitnick, C. L. and Dollr, P. (2014). Microsoft coco: Common objects in context, *ArXiv14050312 Cs* .
- Lu, J., Xiong, C., Parikh, D. and Socher, R. (2017). Knowing when to look: Adaptive attention via a visual sentinel for image captioning, *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* .
- Malik, P. and Baghel, A. S. (2019). Performance enhancement of machine translation evaluation system for english-hindi language pair, *I.J. Modern Education and Computer Science* **2**: 42–49.
- Mao, J., Xu, W., Yang, Y., Wang, J., Huang, Z. and Yuillel, A. (2014). Deep captioning with multimodal recurrent neural networks (m-rnn), *ArXiv14126632 Cs* .
- Mariscal, G., Marban, O. and Fernandez, C. (2010). A survey of data mining and knowledge discovery process models and methodologies, *Knowl. Eng. Rev* **25**(2): 137–166.
- Miyazaki, T. and Shimizu, N. (2016). Cross-lingual image caption generation, *In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics* pp. 1780–1790.
- Nair, J., Krishnan, K. A. and Deetha, R. (2016). An efficient english to hindi machine translation system using hybrid mechanism, *in 2016 International Conference on Advances in Computing, Communications and Informatics (ICACCI)* pp. 2109–2113.
- Papineni, K., Roukos, S., Ward, T. and Zhu, W. (2002). Bleu: A method for automatic evaluation of machine translation, *in Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, Stroudsburg, PA, USA* pp. 311–318.
- Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition, *ArXiv14091556 Cs* .
- Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J. and Wojna, Z. (2016). Rethinking the inception architecture for computer vision, *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA* pp. 2818–2826.

- Tanti, M., Gatt, A. and Camilleri, K. P. (2017). What is the role of recurrent neural networks (rnns) in an image caption generator?, *arXiv:1708.02043 [cs]* .
- Tanti, M., Gatt, A. and Camilleri, K. P. (2018). Where to put the image in an image caption generator, *Nat. Lang. Eng.* **24**(3): 467–489.
- Tsutsui, S. and Crandall, D. (2017). Using artificial tokens to control languages for multilingual image caption generation, *ArXiv170606275 Cs* .
- van Miltenburg, E., Elliott, D. and Vossen, P. (2017). Cross-linguistic differences and similarities in image descriptions, *In Proceedings of the 10th International Conference on Natural Language Generation* .
- Verma, K. and Singh, M. (2018). Hindi handwritten character recognition using convolutional neural network, *International Journal of Computer Sciences and Engineering* **6**(6).
- Vinyals, O., Toshev, A., Bengio, S. and Erhan, D. (2015). Show and tell: A neural image caption generator, *Presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* pp. 3156–3164.
- Vinyals, O., Toshev, A., Bengio, S. and Erhan, D. (2017). Show and tell: Lessons learned from the 2015 mscoco image captioning challenge, *IEEE Trans. Pattern Anal. Mach. Intell.* **39**(4): 652–663.
- Wirth, R. (2000). Crisp-dm: Towards a standard process model for data mining, *in Proceedings of the Fourth International Conference on the Practical Application of Knowledge Discovery and Data Mining* pp. 29–39.
- Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhutdinov, R., Zemel, R. and Bengio, Y. (2015). Show, attend and tell: Neural image caption generation with visual attention, *ArXiv150203044 Cs* .
- Yoshikawa, Y., Shigetoand, Y. and Takeuchi, A. (2017). Stair captions: Constructing a large-scale japanese image caption dataset., *CoRR abs/1705.00823* .
- You, Q., Jin, H., Wang, Z., Fang, C. and Luo, J. (2016). Image captioning with semantic attention, *Presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* pp. 4651–4659.
- Young, P., Lai, A., Hodosh, M. and Hockenmaier, J. (2014). From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions, *Trans. Assoc. Comput. Linguist* pp. 67–78.