Image Caption Generation Using Text Augmentation For Indian Languages

A THESIS

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR DEGREE OF

BACHELOR OF TECHNOLOGY

IN

INFORMATION TECHNOLOGY

By

Mohan Agarwala(IIB2020014)
Sanjay Ram(IIT2020247)
Palen Pushkar(IIT2017042)

UNDER THE SUPERVISION OF

Dr. Naveen Saini

INDIAN INSTITUTE OF INFORMATION TECHNOLOGY, ALLAHABAD

(A UNIVERSITY ESTABLISHED UNDER SEC.3 OF UGC ACT, 1956 VIDE NOTIFICATION NO. F.9-4/99-U.3 DATED 04.08.2000 OF THE GOVT. OF INDIA)

A CENTRE OF EXCELLENCE IN INFORMATION TECHNOLOGY ESTABLISHED BY GOVT. OF INDIA

JULY 2024

CANDIDATE DECLARATION

We, do hereby declare that this thesis titled, **"Image Caption Generation Using Text Augmentation For Indian Languages"** submitted in the qualified perfection for the degree of **Bachelor of Technology in Information Technology** at **Indian Institute of Information Technology, Allahabad** is a record of bonafide work done by me under the able guidance of **Dr. Naveen Saini** and due acknowledgements have been made to all the other material used. This report work was done in full compliance with the requirements and constraints of the prescribed curriculum

Place : Prayagraj
Date: / /

Mohan Agarwala(IIB2020014)

Sanjay Ram(IIt2020247)

Palen Pushkar(IIT2017042)

CERTIFICATE FROM SUPERVISOR

I do hereby recommend that the thesis prepared under my supervision by **Mohan Agarwala(IIB2020014), Sanjay Ram(IIT2020247), Palen Pushkar(IIT2017042)** entitled "**Image Caption Generation Using Text Augmentation For Indian Languages**" be accepted in partial fulfillment for the degree of B.Tech in Information technology for examination.

Date: / /
Place: Prayagraj

Dr. Naveen Saini

Thesis Supervisor

Countersigned by:

.....

Dean(Academics)

CERTIFICATE OF APPROVAL

The foregoing thesis is hereby approved as a creditable study in Information Technology and its allied areas. It is carried out and presented in a satisfactory manner to warrant its acceptance as a prerequisite to the degree for which it has been submitted. It is understood that by this approval the undersigned do not necessarily endorse or approve any statement made, opinion expressed or conclusion drawn therein but the thesis only for the purpose for which it is submitted.

Committee Members for Evaluation of the Thesis for Final Examination:

.....

.....

.....

.....

Acknowledgements

It is my honour and privilege to get an opportunity to study at such a great institute. **"Indian Institute Information Technology –Allahabad"**, where each day was a chance to learn and grow personally and academically and professionally. The wealth of knowledge and experience, it has given us is deep-rooted foundations in our various fields of study.

I would like to express my sincere gratitude to all those who provided me with the resources and guidance to complete my thesis. I am profoundly grateful to **Dr. Naveen Saini** for his expert guidance and continuous encouragement throughout to see that this project rights its target since its commencement to its completion. His instrumental contribution in the proposed work has made it conceivable.

Finally at the end I would like to thank all the faculty members who inspired me to finish my work.

Mohan Agarwala
Sanjay Ram
Palen Pushkar

BTech I.T

ABSTRACT

With applications in a variety of fields, including image retrieval, categorizing and finding users' interest in images, etc., image captioning is a fascinating and difficult issue. It possesses enormous potential to replace the tedious process of creating captions for photos, especially when dealing with big amounts of image data. The limited set of words in the dataset constitute a problem while generating captions and resulting into irregularity and wrong captions.

Deep neural networkbased techniques have seen a lot of success recently in the fields of language production, computer vision, and machine translation. However, as the majority of people in India don't speak English, the country's large and diversified speaking population can benefit from the creation of image captions. Therefore, we translate the models into Hindi and Odia. We employ the cutting edge LLMELECTRA for text augmentation in order to increase the dataset's size and comprehensiveness, resulting in a wider range of outcomes.

Index Terms– Image, Neural Network, Caption, CNN, Feature Extraction, RNN, Hindi, ELECTRA, ODIA

The main contributions made in this thesis are:

- We proposed the use of augmentation of image captions in a dataset (including augmentation using ELECTRA) to improve a solution of the image captioning problem.
- We translated the COCO dataset thus generating captions for Hindi and Odia languages using GoogleTranslator
- We implement the models for Hindi and Odia language to include more Indian languages in the context of image caption generation

Contents

List of Figures

Chapter 1

INTRODUCTION

In computer vision, automatic picture caption generation is an ongoing research area. Because this topic combines two of the key domains of artificial intelligence, computer vision and natural language processing, and has a wide range of practical applications, researchers are drawn to it. Understanding an image is a prerequisite for creating a meaningful word out of it, and object recognition and image classification can help with this [2]. In reality, object detection and image classification are easier tasks than automatic caption synthesis. By looking at photos, people can create captions for them. Finding visual things that are represented in a picture and determining the relationship between those images are inherent human abilities.[1] Caption generation is a skill that is acquired via learning and experience. It is theorized that machines can learn to comprehend the relationships between objects and attain accuracy comparable to humans by using a variety of datasets for training. With the significant advancements in Artificial Intelligence (AI), photos are now being used as input for a variety of functions.In the paper [17], one application of AI has been covered. They identified the face using deep learning techniques. The primary goal of automatic image caption generation is to produce coherent sentences that explain the image's content and the relationships between the items that are identified in the image. These sentences can then be used as recommendations in a variety of applications. It can be applied to a number of natural language processing tasks, in cluding social media recommendation, image indexing, virtual assistants, and visually impaired people [10, 5]. The creation of picture captions can aid machines in comprehending the content of images. It involves more than just finding objects in a picture it also entails figuring out how the objects that have been found relate to one another. Image captioning techniques are divided into three categories by researchers: deep neural networkbased, retrievalbased, and templatebased [16]. In templatebased approaches, attributes, objects, and actions are first identified from the image, and then blank slots in predetermined templates are filled

in. Retrieving an image that resembles the input image is how retrievalbased approaches generate captions. While syntactically accurate captions are produced by these approaches, semantic accuracy and visual specificity cannot be guaranteed. Using a linguistic model, captions are generated once the image has been encoded in deep neural networkbased approaches. It involves more than just finding objects in a picture it also entails figuring out how the objects that have been found relate to one another. Image captioning techniques are divided into three categories by researchers: deep neural networkbased, retrievalbased, and templatebased [16]. In templatebased approaches, at tributes, objects, and actions are first identified from the image, and then blank slots in predetermined templates are filled in. Retrieving an image that resembles the input image is how retrievalbased approaches generate captions. While syntactically accurate captions are produced by these approaches, semantic accuracy and visual specificity cannot be guaranteed. Using a linguistic model, captions are generated once the image has been encoded in deep neural network based approaches. In contrast to the first two approaches, the deep neural networkbased approach might provide better results. We utilized text augmentation techniques on image captions extracted from an MSCOCO dataset. Data augmentation is a commonly employed technique to enhance performance and achieve model stability in many machine learning appli cations. Various language models can be utilized for text augmentation. ELECTRA, established in [14], is currently one of the most successful language models. ELECTRA utilizes the Transformer model, which employs an attention mechanism to acquire contextual relationships among words (or subwords) in a given text. A Transformer is composed of two distinct components—a text input reader called an encoder, and a prediction generator known as a decoder. Unlike directional models that process text input sequentially, the Transformer encoder comprehends the complete sequence of words simultaneously. This attribute enables the model to acquire an understanding of the meaning of a word by considering its entire context

1.1 ENCODER MECHANISM

The encoder component takes an input image and converts it into a fixed-size feature representation. A pre-trained convolutional neural network , is used as the encoder which extracts visual features from the image using convolutional layers and produces a feature vector. CNN uses the convolution of different nodes, here convolution is similar to applying filters. It is suited for grid like structures. The convolution

is applied which reduces the size of the data, and then pooling(taking the average or max of the grid) is used which further reduces the data. and then in the end all the data is put into a linear vector format. The idea is to extract features out of the image, which are tokenised into numbers for machines to understand

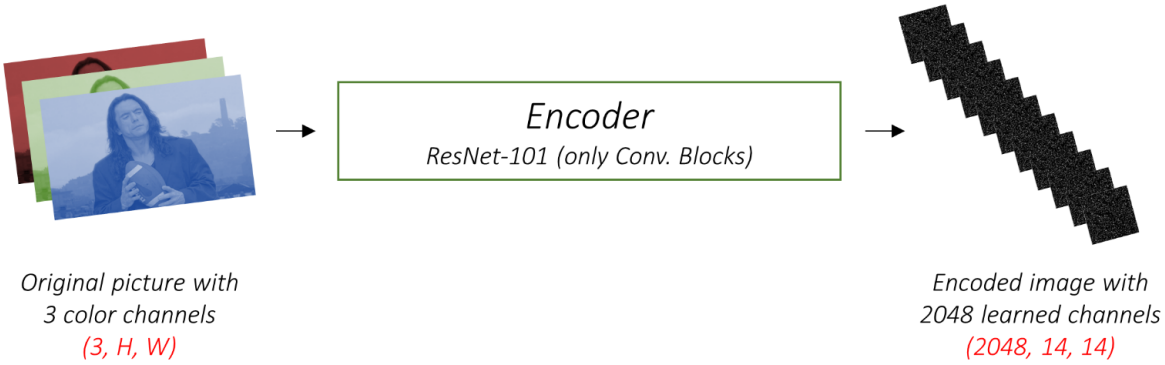


Figure 1.1: Architecture of encoder model.

1.2 DECODER MECHANISM

Decoder Mechanism The decoder component generates a sequence of words that form the caption based on the encoded image features.A recurrent neural network (RNN), is used as the decoder that takes the encoded image features as input and generates one word at a time, conditioning on the previously generated words. At each time step, the decoder predicts the next word in the sequence using a softmax layer over the vocabulary. RNN uses hidden layers to remember the sequence of words like a timeseries. All the captions of the images are put into a dictonary and the RNN will predict the following words from the previous word and feature vector using the dictonary.

1.3 ATTENTION

They enable models to focus on specific parts of the input sequence when making predictions or generating by RNN. There are different types of attention model viz. self attention, scaled dot attention(It takes dot product of queris and key-value pair and applies softmax function), multi head attention (It has multiple attention in parallel and then fused together.)

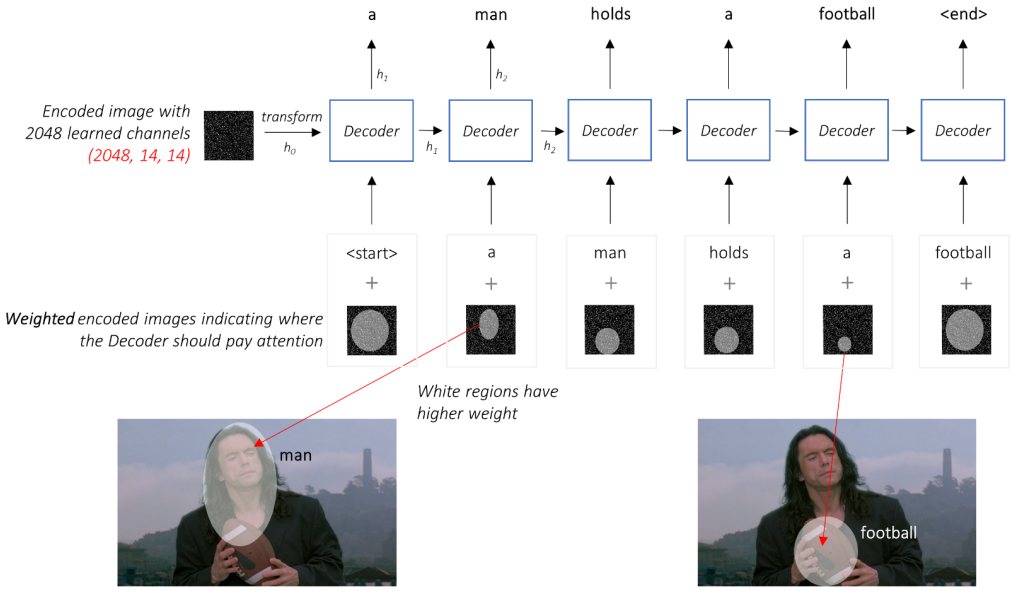


Figure 1.2: Architecture of decoder model.

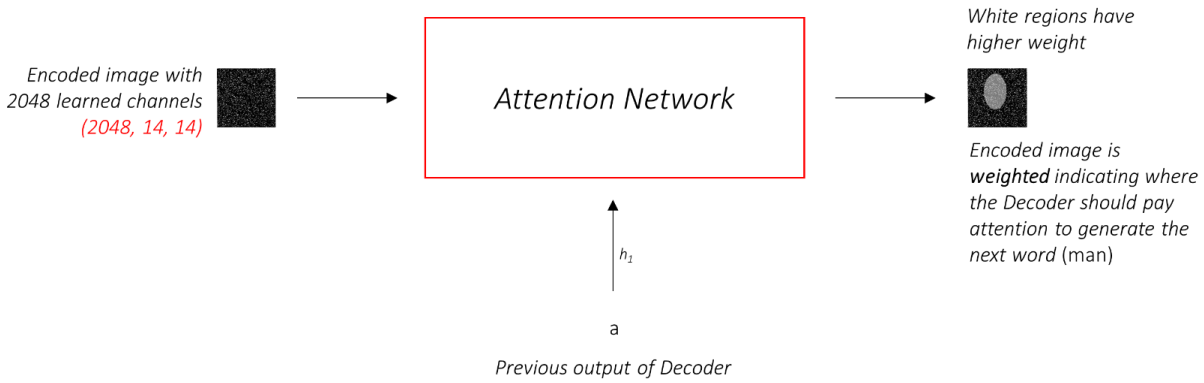


Figure 1.3: Attention Mechanism

1.4 ELECTRA

The idea behind ELECTRA is to teach the model to predict missed words in a sentence. In order to do this, part of the words in the sentence are replaced with a MASK and the task of the model is to predict those words by their context.[2] Hence ELECTRA is pre-trained on a large set of texts in an unsupervised manner having high-quality vector representations of words and a model that can predict words by context and the presence of a connection between sentences. Glove is a vector of words made from twitter and wikipedia data.

1.5 DATASET

Widely used for computer vision and object detection related works, Coco-2014 dataset [4] contains manually labelled captions for each image. The dataset contains

images and the corresponding captions in English language. The dataset has two parts: image directory and annotations section. The image directory contains 80000 images with annotations and 80 object categories. The dataset contains images in JSON format divided into training, testing and validation sets. The annotation section contains other details about the corresponding image such as date of capture, size, imageId and licence number. Each image has 5 captions which are generated manually of variable size.

The dataset is divided into training set (ap- prox. 83K images), validation set (approx. 41K images) and test set (approx. 40K images). A few sample images from the dataset, along with their reference captions in English, are shown . All the images are in jpg format. The average length of the captions is 12 . The resolution of the input images is 256×500 to 500×500

Figure 1.4: Sample image with reference captions.

1. A guy is riding a bike up the side of a hill.
2. A young man bicycles towards the camera and away from beautiful mountains on a clear day.

3. Man on bike in mountains.
4. Man riding a bicycle down a narrow path.
5. Man riding bike on trail.

1.6 EVALUATION METRICES

To evaluate the generated captions, the metrices compare different aspects of the captions such as readability, semantics, grammer, and content. The widely used matrices are discussed below. [5]

The BLEU score, utilized to evaluate machine translation outputs, is determined by comparing the ngrams in the generated text with those in the reference text, tallying the matches. It's a precision-focused measure, with a score above 0.5 generally deemed excellent and below 0.15 indicating significant room for improvement in the model. However, one drawback of BLEU is its disregard for semantic meaning and sentence structure.

To counter BLEU's limitations, METEOR was developed, which computes its score based on unigram recall and precision. METEOR is designed to better align with human judgment, placing greater emphasis on recall than precision.

GLEU, on the other hand, was designed to evaluate sentence fluency and grammatical accuracy. It assesses overlapping ngrams with a set of reference sentences and isn't compatible with error precision/recall measures. At the corpus level, GLEU scores tend to be quite similar to BLEU scores.

ROUGEL is predominantly used for assessing the quality of image captions. It identifies the longest matching sequence of words using the longest common subsequence (LCS) and doesn't require continuous word matches.

Structure of the thesis is organized as follows: CHAPTER 2 covers background work done and model definitions. CHAPTER 3 contains a brief descriptions of problem statement. CHAPTER 4 briefs the proposed methodology. CHAPTER 5 demonstrates experimental results and observations. CHAPTER 6 finally winds up with a conclusion.

Chapter 2

LITERATURE REVIEW

All the methods discussed below were Image Caption Generation has been a budding field and an extensive ammount of work has been done in image captioning and content generation for for image captioning. The beginning phase was characterised by content generation which can be classified as visual, conceptual and geometric. Along with it, vareity of methods have been proposed for image caption generation viz, Template based methods, retreival based methods, and Deep neural network based methods. In template based methods, first attributes, objects and actions are detected from the image and then predefined templates with number of blank slots are filled. In retrieval based methods, caption is generated by retrieving an image that is similar to the input image. These methods generate syntactically correct captions, although image specificity and semantic correctness is not guaranteed. In Deep neural network based methods, first image is encoded and then captions are generated using language model. CNN uses the convolution of different nodes, here convolution is similar to applying filters. It is suited for grid like structures. The convolution is applied which reduces the size of the data, and then pooling(taking the average or max of the grid) is used which further reduces the data. and then in the end all the data is put into a linear vector format. The idea is to extract features out of the image, which are tokenised into numbers for machines to understand. RNN uses hidden layers to remember the sequence of words like a timeseries. All the captions of the images are put into a dictonary and the RNN will predict the following words from the previous word and feature vector using the dictonary. The deep learning based methods often take MSCOCO and Flickr30K dataset as reference.

2.1 Template Based

The template-based method generates captions for objects, activities, and properties identified in the input image by using predefined templates with blank spaces.

The authors proposed filling template slots with expected triplets of visual elements (scene, object, and action). Similar to this, in [10], characteristics and prepositions were identified together with things (like people, automobiles), or elements (like trees, roads), using the Conditional Random Field (CRF) approach. BLEU and ROUGE scores were evaluated using the PASCAL dataset; the highest BLEU score was 0.18, and the corresponding ROUGE score was 0.25.

Another strategy is to carefully combine significant phrases from preexisting captions to create new ones by selecting them, as explained in [9]. BLEU (0.189) and METEOR (0.101) scores were calculated on a test set of 1000 photos from a sample of one million captioned images. Although these techniques are good at producing captions that follow grammar rules, their capacity to create captions with varying lengths is limited because of their heavy reliance on pre-made templates.

In order to tackle this issue, [8] suggested a captioning model for images that is augmented in memory. By encoding prior knowledge through external memory, this model improves the decoder's capacity to produce precise captions. A 3.5 improvement in CIDr was found when evaluating the MSCOCO dataset in comparison to baseline models.

2.2 Retrieval Based

Retrieval-based captioning techniques construct captions for images by grouping together visually comparable images. Following the identification of visually similar photos, these methods search the training dataset for captions for visually comparable images, then use those captions to construct the caption for the query image. The authors of [20] developed a model to find similar images in a big dataset and provide the descriptions of these retrieved images for the query image by utilizing millions of photos and their descriptions.

A density estimation technique was used in [15] to produce captions, and the resulting BLEU score was roughly 0.35. Similar to this, in [16], similar photos were clustered using visual and semantic similarity scores. Then, the images were merged, and the input image's caption was extracted from the captions of similar images in the same cluster.

As mentioned in [10], some researchers developed a ranking-based framework that integrated sentence-based picture captioning to provide captions for every image. Furthermore, a text-based visual attention (TBVA) model for automatically recognizing salient objects was presented by the authors in [18]. The suggested

model was tested using datasets like Flickr30k and MSCOCO.

Additionally, a data-driven method for retrieval-based technique-based image description synthesis was presented in [39]. The study came to the conclusion that the suggested approach for creating picture captions provides effective and pertinent outcomes.These tactics frequently fail to produce image-specific and semantically right statements, even while they produce sentences that are syntactically legitimate and generic.

The BagLSTM approach, along with LSTM variants, was introduced in [7] for the purpose of automatic picture captioning on the MSCOCO dataset. The authors determined that BagLSTM had superior performance compared to other variations based on its CIDEr value. In addition, a text feature extraction method for image captioning was proposed in [17]. This method utilized deep neural networks (DNN) with LSTM and was assessed on the Flickr30k dataset.

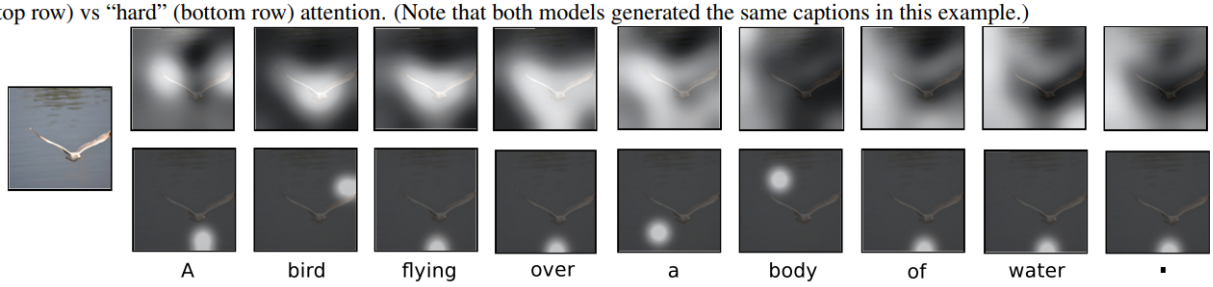
A study introduced a semantic embedding as a global guiding and attention model in [21]. To verify the effectiveness of this technique, experiments were carried out on the Flickr8k, Flickr30k, and MSCOCO datasets. Paper [6] introduced a method for image captioning that combines an RCNN-based top-down and bottom-up approach. This method is improved by employing beam search decoders and explanatory features to reorganize the captions.

In addition, [11] introduced a Reference-based Long Short-Term Memory (RLSTM) technique for generating image captions automatically. This approach utilizes a weighting system to determine the significance of words and images in order to generate appropriate captions. The Flickr30k and MSCOCO datasets were subjected to validation, resulting in a significant 10.37 rise in the CIDr value specifically for the MSCOCO dataset.

2.3 Deep Learning Based

Upon examining the papers on image caption generation using deep learning,some noteworthy findings were investigated. Firstly (RP1), the majority of cutting-edge CNN models are pretrained on the ImageNet dataset, which focuses on single objects rather than entire scenes. As a result, these models tend to produce object-specific outcomes. In order to tackle this issue, the suggested model utilized the VGG16 Hybrid Places1365 model, which had been pre-trained on both the ImageNet and Places datasets. This allowed the model to generate results that were particular to different scenes. Furthermore, numerous articles presented findings based on a lim-

Figure 2. Attention over time. As the model generates each word, its attention changes to reflect the relevant parts of the image. "soft" (top row) vs "hard" (bottom row) attention. (Note that both models generated the same captions in this example.)



ited number of evaluation criteria, including accuracy and BLEU1 score. In order to address this issue, the suggested model was assessed using several metrics such as BLEU, METEOR, ROUGE, and GLEU measurements.

Chapter 3

PROBLEM FORMULATION

3.1 Challenges

We encountered the following challenges and gaps in the research papers discussed above in the field of image caption generation.

- Limited number of captions and contextual data
- Lack of inclusion of Indian Languages

3.2 Solutions

We came up with the following ways to counter the challenges

- Using text augmentation to enrich the dataset
- Inclusion of Hindi and Odia Language

Chapter 4

PROPOSED METHODOLOGY

4.1 Dataset

Odia And Hindi Language : MSCOCO-2014 dataset was taken as the base dataset whose annotations were translated to Hindi and Odia. The vocab was built by taking the words from the captions and removing the less frequent words. Further ELECTRA and GLOVE were added to enhance the vocab thus making it more comprehensive.

1) "प्रकृति की गोद में उद्योग की छाया"
2) "हरियाली से घिरा औद्योगिक ढांचा"
3) "शांति और संघर्ष का संगम"
4) "विकास की ओर अग्रसर प्रकृति"
5) "आधुनिकता की ओर बढ़ते कदम"

1) "ସଡ଼କ ପାର୍ଶ୍ୱରେ ମିଳନ"
2) "ଖାଦ୍ୟ ଗାଡ଼ି ଆଉ ଆଡ଼ମ୍ବର"
3) "ବିକଳେର ଆନନ୍ଦ"
4) "ସାମାଜିକ ମଳେ।"
5) "ଏକାଠି ଖାଇବା।"

4.2 Feature Extraction

The encoder is used for visual feature extraction of a picture. CNN are commonly used as encoders due to its advantage in dealing with grid like structures.

4.3 Sentence Generation

This component uses RNN based model to generated sentences. The output of the feature extraction model is connected to the RNN which uses the vocabulary along with the feature vectors to generated new words. In other words, it uses the previous word along with the feature vector to generate the next word from the vocabulary. [3]

Vanishing gradient occurs during the training where the gradients that are used to update the network become extremely small as they are backpropogated from the output layers to the earlier layers, hence LSTM is used to counter this issue. Also, Instead of choosing the word with highest probability, the model chooses the sequence that has the highest overall score from a basket of candidate sequences, which is also called as Beam-searching.

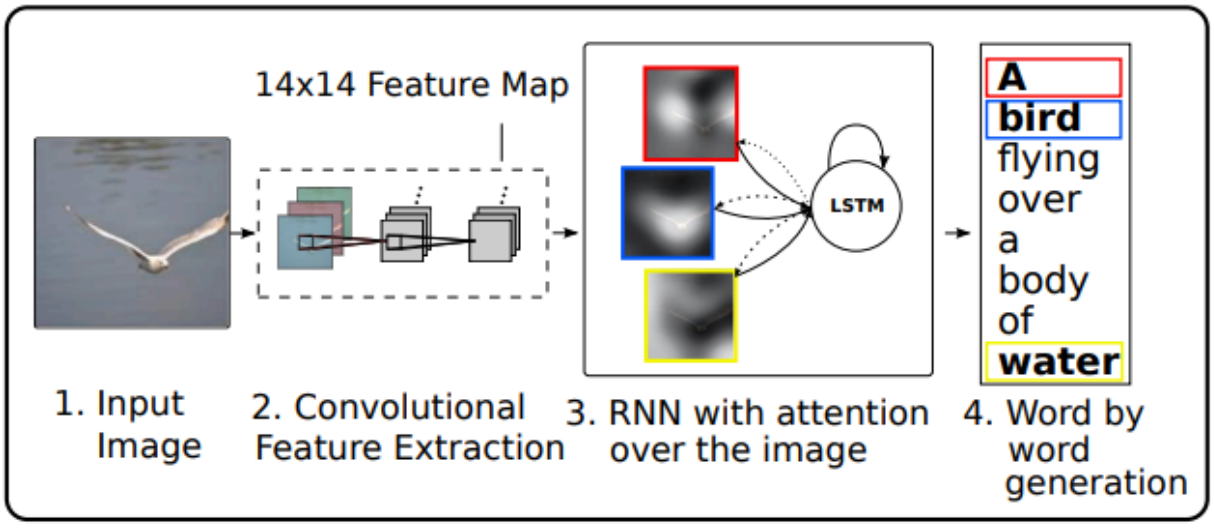


Figure 4.1: Architecture of proposed model.

Chapter 5

Result and Analysis

The expected characteristics of the project was to generate captions for Hindi and Odia language and train the model in those languages. Firstly the captions were translated into Hindi and Odia using translator and some of them were manually checked to see the output generation. After satisfactory results from the translator, the captions were fed into the model for vocabulary building. With the help of ELECTRA and Glove, to be used as language dict There are two desirable characteristics expected from the generated caption after the model has been trained. Firstly, it should correlate with all the objects present in the image. Secondly, it should be useful and understandable to human beings.

The provided table presents performance metrics, including BLEU, METEOR, ROUGE-L, and CIDEr scores, for caption generation in English, Hindi, and Odia, both with and without augmentation. Let's analyze the conclusions and explanations based on the data:

5.1 Performance Disparity:

English captions consistently outperform both Hindi and Odia captions across all metrics. For instance, in terms of BLEU score, English captions score 38.3 without augmentation and 37.8 with augmentation, while Hindi captions score 30.5 and 31.2, and Odia captions score 23.7 and 24.5, respectively. This indicates that the models trained on English data are more effective at generating captions that align closely with human-generated ones compared to Hindi and Odia.

5.2 Impact of Augmentation:

Augmentation techniques show a minor improvement in performance for all languages. However, the differences in scores between captions with and without aug-

mentation are relatively small. For example, the BLEU score for English captions changes from 38.3 to 37.8 with augmentation, indicating a slight decrease in performance. Similar marginal changes are observed for Hindi and Odia captions.

5.3 Challenges with Odia:

Odia captions consistently exhibit the lowest performance scores among the three languages across all metrics. For instance, the BLEU score for Odia captions is substantially lower compared to English and Hindi, indicating that the models struggle to generate captions that match human references. This disparity suggests that the models trained on Odia data may face challenges related to vocabulary, syntax, or dataset size, leading to poorer performance.

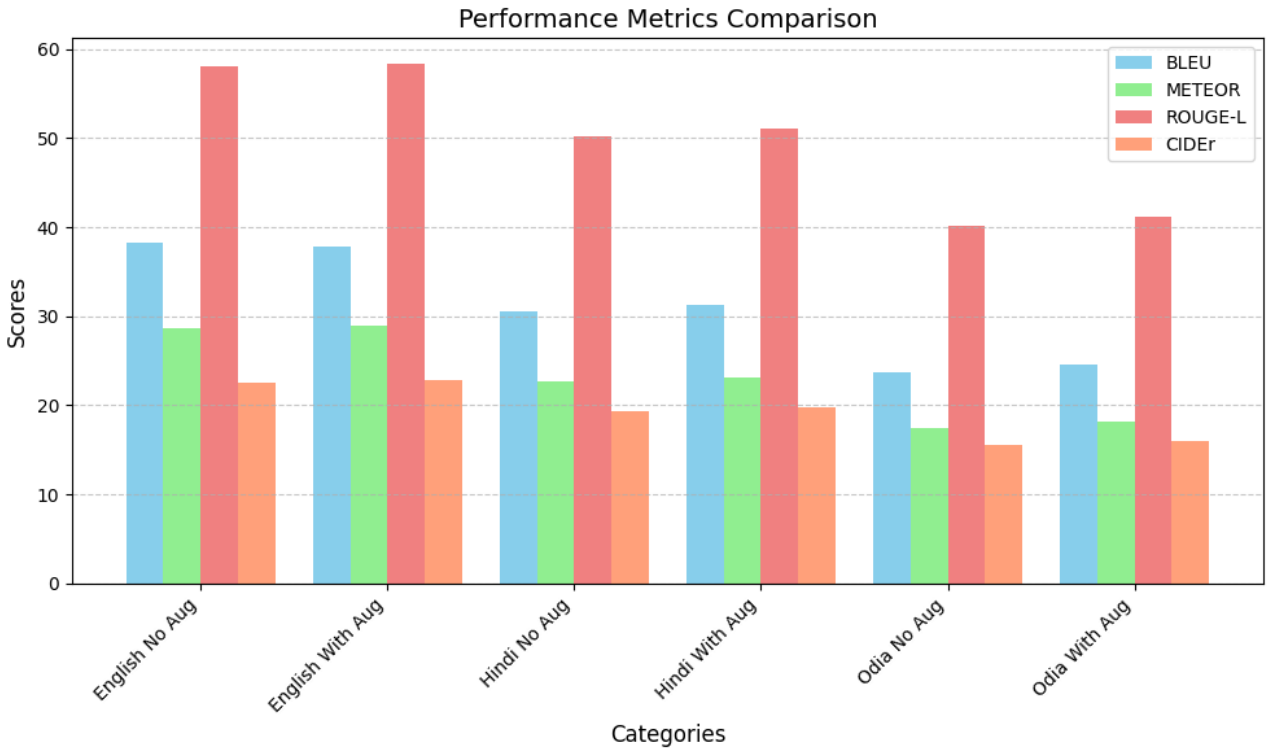
The superior performance of English captions can be attributed to the extensive research and resources available for the English language in the field of natural language processing. Pre-trained models, large-scale datasets, and advanced techniques contribute to the higher performance observed for English captions. While Hindi benefits from pre-trained models like Electra, the performance still falls short of English captions. This suggests that further research and development efforts are needed to enhance the effectiveness of NLP models for languages other than English. Also, The generated captions were not following all the rules of grammar. High training time was observed and therefore, there is some scope for improving the efficiency of the training process. Last, but not the least, the generated captions appear to be less descriptive in comparison to the reference captions. Also, a few failure cases were noted during the experiments.

Performance Metrics Comparison

	BLEU	**METEOR**	**ROUGE-L**	**CIDEr**
English Captions (No Augmentation)	38.3	28.6	58.0	22.6
English Captions With Augmentation	37.8	28.9	58.3	22.8
Hindi Captions (No Augmentation)	30.5	22.7	50.2	19.4
Hindi Captions With Augmentation	31.2	23.1	51.0	19.8
Odia Captions (No Augmentation)	23.7	17.5	40.1	15.6
Odia Captions With Augmentation	24.5	18.1	41.2	16.0

4.2 Sample result caption

In order to demonstrate the validity of the proposed model, we show the generated captions for ten sample images from the dataset along with their reference captions



(ground-truth) in Figure. The result on these random images further strengthens the claimed effectiveness of the model.

 Reference Captions: 1) "प्रकृति की गोद में उद्योग की छाया" 2) "हरियाली से घिरा औद्योगिक ढांचा" 3) "शांति और संघर्ष का संगम" 4) "विकास की ओर अग्रसर प्रकृति" 5) "आधुनिकता की ओर बढ़ते कदमें" Generated Caption: "उद्योग और प्रकृति का सामंजस्य" BLEU Score: 0.35	 Reference Captions: 1) "हरी भरी ब्रोकोली की खूबसूरती" 2) "प्रकृति की कलाकारी: ब्रोकोली का जाल" 3) "ब्रोकोली: हरियाली का चमत्कार" 4) "स्वास्थ्य का खजाना: ताजी ब्रोकोली" 5) "बागवानी का गौरव: ब्रोकोली" Generated Caption: "ब्रोकोली की रंगत में छिपा स्वाद" BLEU Score: 0.61
 Reference Captions: 1) "विंटेज बस का शानदार सफर" 2) "यात्रा की पुरानी यादें: बस संख्या 37" 3) "ब्लैकवुड और न्यू ट्रेडेगर की ओर" 4) "IBT की धरोहर: नंबर 37 बस" 5) "अतीत की ओर एक झलक" Generated Caption: "समय के पन्नों से: विंटेज बस" BLEU Score: 0.26	 Reference Captions: 1) "धारियों का खेल: ज़ेबरा की अनूठी छटा" 2) "प्रकृति की कलाकृति: ज़ेबरा की पट्टियाँ" 3) "जंगल की सुंदरता: ज़ेबरा के रंग" 4) "वन्य जीवन की विविधता: ज़ेबरा" 5) "ज़ेबरा की धारियाँ: प्रकृति का चमत्कार" Generated Caption: "काली और सफेद पट्टियों का संगम" BLEU Score: 0.24
 Reference Captions: 1) "दोस्तों का साथ" 2) "सड़क किनारे मस्ती" 3) "खाने की गाड़ी के पास भीड़" 4) "शाम की चहल-पहल" 5) "बाजार का नजारा" Generated Caption: "मिल-जुल कर खाना" BLEU Score: 0.35	 Reference Captions: 1) "ସଡ଼କ ପାର୍ଶ୍ୱରେ ମିଳନ" 2) "ଖାଦ୍ୟ ଗାଡ଼ି ଆଉ ଆଡ୍‌ମ୍ବର" 3) "ବିକଳେର ଆନନ୍ଦ" 4) "ସାମାଜିକ ମଜୋ" 5) "ଏକାଠି ଖାଇବା" Generated Caption: "ବଜାର ଦୃଶ୍ୟ" BLEU Score: 0.23
 Reference Captions: 1) "ରଳେ ପଥର ଆଧୁନିକ ଯାତ୍ରା" 2) "ବିଦ୍ୟୁତ୍ ଲାଇନ୍ ଉପରେ ଚାଲୁଛି ଟ୍ରେନ୍" 3) "ପାଟି ଉପରେ ହଳଦିଆ ଟ୍ରେନ୍" 4) "ଯାତ୍ରାର ନୂଆ ମୁହାଁ" 5) "ଟ୍ରେନ୍ ଏବଂ ତାର ଜାଲିକା" Generated Caption: "ଗତିଶୀଳ ଟ୍ରେନ୍, ଗତିଶୀଳ ଜୀବନ" BLEU Score: 0.29	 Reference Captions: 1) ପିଜ୍ଜା ସହ ଆନନ୍ଦ 2) ଶିଶୁ ଉତ୍ସବ 3) ଖୁସିର କ୍ଷଣ 4) ମଜାଦାର ଭୋଜନ 5) ପାର୍ଟିର ମଜା Generated Caption: ସଙ୍ଗରେ ଖାଇବା BLEU Score: 0.26
 Reference Captions: 1) समयर साक्षी 2) ঐতিহাসিক ঘড়ি 3) সংস্কৃতির স্মৃতি 4) বিকাশের চিহ্ন 5) বিশ্বধরার সাক্ষী Generated Caption: কলার প্রকাশ BLEU Score: 0.29	 Reference Captions: 1) ରଙ୍ଗୀନ ଏକାକୀ 2) ବିବରଣ ପୃଷ୍ଠପଟରେ ଉଜ୍ଜ୍ୱଳ 3) ସଡ଼କ ପାର୍ଶ୍ୱର ଆକର୍ଷଣ 4) ନୀଳ ମୁଣ୍ଡର ହଳଦିଆ ସାଥୀ 5) ଅନ୍ୟାନ୍ୟ ଉଜ୍ଜ୍ୱଳତା Generated Caption: ହଳଦିଆ ଅଗ୍ନିଶାମକ BLEU Score: 0.23

Table I
Images with Reference and Generated Captions

Chapter 6

Conclusion And Future Work

6.1 Conclusion

The Project proposed the integration of text augmentation technique to expand the dataset for Hindi and Odia language and used the state of the art models for caption generation. Which can be summarised as follows:

- Integration of ELECTRA
- Dataset generation for Odia language
- Implementation of state of the art models for Hindi and Odia languages

More Indian Languages can be used for Image Caption generation which will benifit the people. Further, attention based models may be employed for strengthening the robustness of the models. Although the model was able to generate captions for Hindi and Odia using the model, Further research and development efforts are needed to enhance the effectiveness of NLP models for languages other than English. Due to lack of embeddings for ODIA and Hindi languages, the models were not able to perform well and give good results. The evaluation matrix were using English language structure hence did not give good results.

6.2 Future Work

- Use for Video Caption generation
- Inclusion of more Indian languages
- Tuning of models for Odia
- Evaluation matrix tuning

- Vocabulary building for Odia Language
- Grammer improvements in generated captions

References

References

[1] Xu, Kelvin, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. "Show, attend and tell: Neural image caption generation with visual attention." In International conference on machine learning, pp. 2048-2057. PMLR, 2015.

[2] Atliha, Viktar, and Dmitrij Šešok. "Text augmentation using BERT for image captioning." Applied Sciences 10, no. 17 (2020): 5978.

[3] Karpathy, Andrej, and Li Fei-Fei. "Deep visual-semantic alignments for generating image descriptions." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3128-3137. 2015.

[4] Lin, Tsung-Yi, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. "Microsoft coco: Common objects in context." In Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13, pp. 740-755. Springer International Publishing, 2014.

[5] https://en.wikipedia.org/wiki/BLEU

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

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

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

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

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

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

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

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

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

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

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

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

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

[14] Ali Farhadi, Mohsen Hejrati, Mohammad Amin Sadeghi, Peter Young, Cyrus Rashtchian, Julia Hockenmaier, and David Forsyth. Every picture tells a

story: Generating sentences from images. In European conference on computer vision, pages 15-29. Springer, 2010.

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