**REMOTE SENSING IMAGE CAPTIONING**

**PHASE-2: PROJECT REPORT**

***Submitted By***

**S.CHANDEESH KUMAR (2015506064)**

**P.NANDHINI (2015506025)**

**S.BADRI NARAYAN (2015506062)**

***in partial fulfillment for the award of the degree***

***of***

**BACHELOR OF TECHNOLOGY**

***IN***

**INFORMATION TECHNOLOGY**

****

**MADRAS INSTITUTE OF TECHNOLOGY**

**ANNA UNIVERSITY, CHENNAI-600 044**

**NOVEMBER 2018**

**BONAFIDE CERTIFICATE**

Certified that this project report **“REMOTE SENSING IMAGE CAPTIONING”** is the bonafide work of **CHANDEESH KUMAR S (2015506064), NANDHINI P (2015506025), BADRI NARAYAN S (2015506062)** who carried out the project work under my supervision. Certified further that to the best of my knowledge the work reported here in does not form in part or full of any other thesis or dissertation on the basis of which a degree or award was conferred on an earlier occasion on this or any other candidate.

|  |  |
| --- | --- |
| **SIGNATURE** | **SIGNATURE** |
| **DR. DHANAJAY KUMAR,** | **Mrs M. HEMALATHA,** |
| **HEAD OF THE DEPARTMENT,** | **SUPERVISOR,** |
| PROFESSOR, | ASSISTANT PROFESSOR, |
| Department of Information Technology, | Department of Information Technology, |
| Madras Institute of Technology, | Madras Institute of Technology, |
| Anna University, Chromepet, | Anna University, Chromepet, |
| Chennai – 600 044 | Chennai – 600 044. |

**ACKNOWLEDGEMENT**

The satisfaction that accompanies the successful completion of our task would be incomplete without mentioning the names of people who made it possible, whose constant guidance and encouragement crowns all efforts with success.

We express our gratitude and sincere thanks to our respected Dean, **Dr.T.Thyagarajan**, Anna University, MIT Campus for being a source of great inspiration to us.

Our heartfelt thanks to **Dr.Dhanajay Kumar,** Head of the Department, Department of Information Technology, Madras Institute of Technology, Anna University, for the prompt and limitless help in providing the excellent computing facilities to do the project and to prepare the thesis.

We express our profound sense of gratitude to our project guide **Mrs.M.Hemalatha ,** Department of Information Technology, for her invaluable support, guidance and encouragement for the successful completion of this project. Her vision and spirit has always inspired and enlightened us.

We express our sincere thanks to the project committee members **Dr.D.Dhananjay Kumar, Dr.J.DhaliaSweetlin**and **Dr.S.UmaMaheswari,** Department of Information Technology, Madras Institute of Technology, Anna University, Chennai, for their valuable guidance and technical support.

**CHANDEESH KUMAR S (2015506064)**

**NANDHINI P (2015506025)**

**BADRI NARAYAN S (2015506062)**

|  |  |  |
| --- | --- | --- |
|  | **TABLE OF CONTENTS** |  |
| **CHAPTER NUMBER** | **TITLE** | **PAGE NUMBER** |
|  | **ABSTRACT** | **i** |
|  | **LIST OF FIGURES** | **ii** |
|  | **LIST OF ABBREVIATIONS** | **iii** |
|  | **LIST OF TABLES** | **iv** |
| **1** | **INTRODUCTION**   * 1. Remote Sensing   2. Image Captioning   3. Image Processing   4. Text Processing | 1 |
| **2** | **LITERATURE SURVEY**  2.1 Natural Image Captioning  2.1.1 Handcrafted Features  2.1.2 Neural Network based Models  2.2 Remote Sensing Image Captioning | **4**  4  4  4  8 |
|  | 2.3 Attention | 9 |
|  | 2.4 Long Short-Term Memory | 10 |
|  | 2.5 Resnet | 11 |
| **3** | **PROPOSED SYSTEM** | **12** |
|  | 3.1 Architecture | 12 |
|  | 3.2 Attention based LSTM | 16 |
|  | 3.3 Image classification | 17 |
|  | 3.4 Domain-Driven method | 17 |
|  | **IMPLEMENTATION DETAILS** | **18** |
| **4** | 4.1 Dataset  4.1.1 UCM Captions Dataset  4.1.2 UAVIC Dataset  4.2 LSTM | 18  18  19 |
|  | 4.3 Bidirectional LSTM | 20 |
|  | 4.4 Word Embeddings | 20 |
| **5** | **RESULTS AND DISCUSSION** | **22** |
|  | 5.1 Experimental Results  5.1.1 BLEU score  5.2 Comparison with other results  5.3 Ground Truth Comparison | 22  22  22  25 |
| **6** | **CONCLUSION AND FUTURE WORK** | **29** |
|  | 6.1 Conclusion and Future work | 29 |
| **7** | **REFERENCES** | **30** |
|  |  |  |

**ABSTRACT**

Image Captioning is the method of generating a sentence for an image, that serves as a description of the image. The sentences consist of words denoting the various objects in the image and the relationship between them. Remote sensing Image Captioning is a special case of Image Captioning which solves the difficulties in training a model when it comes to remote sensing images. Issues may arise due to translation, rotation and viewpoint of images and maintaining semantic consistency in the generated sentence. To represent information present in a remote sensing image, the problem of image captioning aims to generate comprehensive sentences that summarizes the primary content present in the image. This method of describing a remote sensing scene in the form of sentences plays a vital role in numerous fields, such as image retrieval, scene classification, and as a vision companion.

**LIST OF FIGURES**

|  |  |  |
| --- | --- | --- |
| **FIGURE NUMBER** | **TITLE** | **PAGE NUMBER** |
| 3.1 | Architecture | 9 |
| 3.2 | LSTM structure | 11 |
| 4.1 | Dataset Example | 14 |
| 5.1 | Inception V3 | 20 |
| 5.2 | Inception V3 with LSTM | 21 |
| 5.3 | ResNet152 with LSTM | 21 |
| 5.4 | ResNet152 with attention based  B-LSTM | 22 |

**LIST OF ABBREVIATIONS**

LSTM - Long Short Term Memory

CNN - Convolutional Neural Network

RNN - Recurrent Neural Network

BRNN - Bidirectional Recurrent Neural Network

GLU - Gated Linear Unit

ReLu - Rectified Linear Unit

FCN - Fully Connected Networks

BLEU - Bilingual Evaluation Understudy

CBOW - Continuous Bag Of Words

B-LSTM - Bi-Directional Long Short Term Memory

**LIST OF TABLES**

|  |  |  |
| --- | --- | --- |
| **TABLE**  **NUMBER** | **TITLE** | **PAGE**  **NUMBER** |
| 3.1 | Experiment models | 12 |
| 5.1 | Results of handcrafted feature based methods | 18 |
| 5.2 | Results on base paper | 18 |
| 5.3 | Results on reference paper | 19 |
| 5.4 | Comparison of results with the baseline model | 19 |
| 5.5 | Comparison of results from different combinations of dataset | 21 |

**CHAPTER - 1**

**INTRODUCTION**

Information description was initially started with the help of symbols called alphabets grouped into words and sentences. It is possible to describe more information with the help of an image. Image has a lot of information to be derived, that there is always a scope for improvement in describing an image. Image captioning is the method of generating caption for a given image by connecting and finding various relationships between the objects present in the image and providing simple understanding in form of one or two sentences.

**1.1 Remote Sensing**

Remote sensing is the application of using satellites or drones to capture images of earth’s surface without actually making any physical contact with the site. This type of images has their application in the assessing the quality of a land - if it is fertile, infertile or barren, weather conditions, traffic/crowd detection, classifying the type of activity taking place at a location- if it is an illicit activity or allowable activity

**1.2 Image Captioning**

Image captioning, in general terms, refers to the process of creating captions in the form of descriptive sentences for a given image. The captions generated found its usage in intelligence systems and computer vision. Remote Sensing image captioning is a specialized application narrowed by the type of images used. This field will find specific applications as mentioned. The output can be described in the formed of well-developed sentences to give crisp details about land in land quality assessment, about the type of crowd in crowd detection, the type of vehicle in traffic detection and the type of activity taking place in the particular area. The sentence will describe the most important aspects of the image using the terms to describe the objects and relationship between them. The image captioning has always been considered as a combination of image processing and text processing model. The most commonly recognized complication of Image Captioning is the process of combining both image and text worlds.

**1.3 Image Processing**

An image is seen by a computer as a 3-layer stacked 2D array with 3 layers signifying the colors RGB and the values of arrays being the pixel values. Earlier models of image processing involved detecting the objects present in the scene and the type of scene present in the image. The methods include extracting features or useful information like primary objects by interpreting their frequency domain, histogram or using mathematical functions to identify significant caption useful portions of a given image. Later, with the improvement of technologies leading to neural networks, Image Processing models moved towards CNNs which provided automatic feature extraction. The CNNs also solved the problem of rotation ambiguity posed by remote sensing images since CNN is invariant to translations.

**1.4 Text Processing**

The problem of Image Captioning involves processing of texts to generate captions. The problem has evolved to become a mathematical problem since the image and text is converted to numbers to be processed by the models. There are different methods for converting words to integers and processing the words by processing the integers. In general, the image captioning problem is the problem of generating a word, given picture and previous word. The problem is reduced to finding the highest probable word in an array of size of the vocabulary. This can be achieved with the help of RNNs, LSTM and text processing libraries like word2vec.

In the development of Artificial Intelligence, the technologies for image captioning find its application in robotic vision, computer vision, etc... Computer vision refers to the processes and functions of a robot machine to process what it sees. Computers view the world in the form of images and it is processed for it to experience human-like vision. The mathematical processing helps to convert the image data into a form that is describable by a computer and similar processing systems. The text processing is the process of manipulating and interpreting the natural language used by humans in their communication. This step allows for easier processing by the computer systems, reducing the overheads created by alphabet character. The captioning research will have some of its good results in real time processing of an area for quality assessment and deploying such applications to naive users who have no knowledge about the domain. It can also help persons with limited abilities to assess an area and will prove useful as an aid to people having visual difficulties.

**CHAPTER - 2**

**LITERATURE SURVEY**

**2.1 Natural Image Captioning**

**2.1.1 Handcrafted Features**

The retrieved-based method generated sentences based on the retrieval result [1]. These methods search for similar images with the query image firstly, and then generate sentences based on the sentences which describe same kind images in datasets. In this case, the grammatical structure of the generated sentences is very poor, sometimes even unreadable. This is because the quality of generated sentences is closely dependent on the result of image retrieval. The retrieved result would be embarrassed when the content of the image is very different from the pictures in the database which are retrieved. The second kind of method is based on object detection [2], [3]. Object detection-based methods detect the objects in an image and then model the relation between the detected objects. The sentence is generated lastly by a sentence generating model using the information describing the relations of objects in image. The performance of the relation model is vital to the generated sentence. And the result of the object detection also has tremendous influence on the finally generated sentences.

**2.1.2 Neural Network based Models**

In [4] they propose a model that generates descriptions for images and their regions. Datasets consists of images and their sentence descriptions. This model aligns Convolutional Neural Networks (CNN) over image regions and bidirectional Recurrent Neural Networks (RNN) over sentences through multimodal embedding. The presented approach aligns the sentence snippets to the visual regions that they describe through a model. Then treating those correspondences as training data for a second model known as Multimodal RNN which learns to generate the captions for the images. Multimodal RNN defines a probability distribution of the next word given the current word and context from the previous time steps. Limitations in this model is that this model can only generate description of only one input array of pixels at a fixed resolution. Secondly the RNN receives information only through additive interactions which are less expressive than the multiplicative interactions which is a more complicated one.

In [5] they propose a model that automatically creates description for images using natural languages. It requires both image processing and as well as natural language processing. The proposed approach uses Show & Tell model. This model uses recent advancement in image recognition and neural machine translation for image captioning. It uses the combination of Inception-v3 and Long Short-Term Memory (LSTM) cell, where LSTM cell provides language modeling capability and Inception-v3 provides object recognition capability. Inception-v3 model assigns probability of all possible objects in image and convert into word vector. This word vector is provided as an input to the LSTM cells which will then form sentence from this word vector. RNN encodes the variable length input to fixed dimensional vectors. Then these vector representations are decoded to obtain the desired output sequence. Instead of using text as an input to encoder, this model uses image as an input. This image is then converted to a word vector and then translated to caption using RNN as decoder. By implementing Show and Tell model, they can able to generate moderately comparable captions when compared to human generated captions.

In [6] they developed a convolutional image captioning technique. They analyzed the CNN and RNN approaches and concluded that CNN produces more entropy in the output probability distribution. The CNN gives better word prediction accuracy and does not suffer as much from vanishing gradient problem. The drawbacks in RNN-based techniques are, the training process are sequential for a particular image-caption pair. So, the output at nth time step has true dependency on the output at n- time step. Secondly RNN produces lower classification accuracy and still suffer from vanishing gradient problem. The proposed approach works under four layers. First is Input Embedding layer, where the input words are embedded to 512-dimensional vectors. This embedding is concatenated to the image embedding layer and provided as input to the feed-forward CNN module. The second layer is the Image Embedding Layer, where the image features are obtained from the fc7 layer of the VGG16 network. They applied dropout, ReLU nth fc7 to obtain a 512-dimensional embedding. Third is CNN module, it operates on the combined input and image embedding vector. In addition, the used gated linear unit (GLU) activations for four conv layers. The feature dimension after convolution layer and GLU is 512.The fourth is the Classification layer, where the 512-dimensional vectors are encoded into 256-dimensional representation per word. Then the vectors are up-sampled via fully connected layer, and passed through a soft-max to obtain the output word probabilities.

In addition to CNN, attention mechanism was used. They computed separate attention parameters and a separate attended vector for every word. To obtain the attended vector they predicted attention parameters over VGG16 max-pooled conv-5 features. Captions are generated by choosing the word with maximum probability at each step. Instead of sampling the maximum probability words, they also performed beam search with different beam sizes. They performed beam search for both LSTM and CNN methods. With beam size of 3 they got a better score than LSTM. This method does better on all BLEU metrics.

In [7] they developed a CNN+CNN framework for image captioning. They proposed a framework that only employs convolutional neural networks (CNNs) to generate captions. CNN shows relatively strong abilities to tackle very long sequences. The proposed model is compared with and without hierarchical attention. There are four modules in their framework. First module is Vision Module, it is CNN without fully connected layer. They used VGG-16 as the CNN for the vision module. The second one is Language Module which is to model sentences. It is based on CNN without pooling. CNN use kernels and stack multiple layers to model the context. Third one is Attention module which connects the vision module with the language module. To predict different words, different objects in the images should be attended and input into the Prediction module which is the fourth module. It is one hidden layer neural network. The output of this module is the prediction probability of the next word.

In addition, a Hierarchical attention module is used. The CNN+CNN module extracts features from top-level of the language model. In hierarchical attention module, attention vectors are computed at each level of the language model and fed into the next level. In contrast to RNN-based models that calculate attention maps in a left-right manner, whereas this attention maps are calculated in a bottom-up manner which prevents the side way connections in the same layer. This allows our model to be trained in parallel over all words in the sentence, rather than word by word. This approach has shown that the ability of CNN-based framework is competitive to LSTM-based models, but can be trained faster. This model pays attention to corresponding areas in the images in a meaningful way.

In [9] they perform dense captioning for images. They developed an architecture that jointly localizes regions of interest and the describes each with natural language. They propose a Fully Convolutional Localization Network (FCLN) to perform localization and description task. FCLN architecture is based on recent CNN-RNN models developed for image captioning but includes a localization layer. Experiments were done on Visual Genome dataset, which comprises of 94,313 images and 4,100,413 region-grounded captions.

The input image is given to a Convolutional network which is VGG-16 architecture. The output of this network is a set of uniformly sampled image locations, which inputs to the localization layer. The localization layer identifies spatial regions of interest and smoothly extracts a fixed-sized representation from each region. It outputs region coordinates, region scores and region features. Recognition network is a fully-connected neural network that processes region features from the localization network. This network produces a code of a fixed dimension which is passed to RNN language model. Region codes were used to condition the RNN language model. Existing datasets will provide only full image captions, but does not provide individual region captions. Dense captioning aims at providing captions for individual regions. Dataset splits into 5000 each for validation and testing and the rest for training. Evaluation metrics includes METEOR score for language and Average Precision (AP) for localization. In addition, this model supports image retrieval using natural language queries, and localizes these queries in retrieved images.

**2.2 Remote Sensing Image Captioning**

Traditional handcrafted features and simple CNN based methods cannot be plainly used for remote image processing applications. The remote sensing images have special properties that differentiate it from a typical image present in popular image datasets. There are certain ambiguities for objects present in the image in terms of scaling, translation and rotation. There is always a possibility for a occlusion to the objects by natural structures like tree cover, and artificial structures like flyovers. Therefore, special processing is required for getting better results in remote sensing-based applications. Only a few studies are done on remote sensing image captioning. One such study [8] tried to model humanlike language description for a remote sensing image. Humans describes an image in levels of key objects, environment and landscape. The method [8] uses 3 stages to understand a given image in terms of key-instance, environment-elements and landscape analysis by using a neural network based on CNN called Fully Connected Networks (FCNs). FCN is similar to CNN having Convolutional, Pooling and activation layers, but having a loss layer that acts as a decision layer in the end as opposed to a fully connected layer in CNN. FCN outputs label maps instead of class labels (as done by CNN). One primary advantage of using FCNs is it allows inputting of arbitrary size images, allowing high resolution images. The language modeling is done by representing the vocabulary in triplets denoting {element, attribute, relationship}. The generated sentences are not descriptive in terms of remote sensing-based information and are erroneous in certain cases in which the image was deceiving in terms of texture, patterns. It is required in such cases to be aware of contextual information.

**2.3 Attention**

In [12] they proposed an attention-based model that automatically learns to describe the content of images. The proposed model is an encoder-decoder approach combined with attention mechanism. For encoding part, they used a convolutional neural network (CNN) to extract a feature vector. For decoding, they used Long Short-Term Memory (LSTM) which produces a caption by generating one word at every time step conditioned by the previous hidden state and the previously generated words. Attention tells where exactly to look when the neural network is trying to predict parts of a sequence. They discussed about hard and soft attention. Soft attention is deterministic, whereas hard attention is stochastic. An attention map is generated whenever the model infers a new word in the caption. Soft attention multiplies this attention map over the image feature map (obtained from CNN) and sums it up. Since hard attention is not deterministic, it samples one or two features on the feature map, according to the probability density function. The focusing region is done by random sampling. However, the expected region may have higher chance to be focused on than the others. Experiments were done on three benchmark datasets, Flickr8k, Flickr30k and MSCOCO. BLEU and METEOR are the evaluation metrics.

**2.4 Long Short-Term Memory**

In [13] they proposed a CNN-LSTM model. Most of the prior research worked on encoder-decoder framework as a black box, here they explored the effect of emitted words on hidden states in the LSTM. Initially the image was given to a deep convolutional neural network (CNN) which gives the vector representation of the image. The output of CNN is then fed to a LSTM network for decoding. As words are emitted, the current predicted sentence is fed back into the LSTM. During this process the hidden state information is recorded. The recorded information shows that semantically similar words results in similar movements, whereas different words diverges. They added more layers to the decoder LSTM network, which resulted in improvements in the evaluation metrices. Experiments were done on MSCOCO dataset, each image in the dataset holds at least five captions of varying length. Drawbacks in this approach is that due to lack of attention partial errors tend to occur. This approach achieved a BLEU-4 score of 3.3 and 3.8 CIDEr score.

**2.5 ResNet**

In [14] they developed a residual learning framework to ease the training part of the neural networks which are deep. They performed several experiments with plain network and residual network. Experiments were done on ImageNet 2012 classification dataset. Initially [14] evaluated 18-layer and 34-layer plain net and found that 34-layer net has higher validation error than the 18-layer net. Then they evaluated 18-layer and 34-layer residual nets (ResNets). Major observations are noted. First, the 34-layer ResNet exhibits lower training error. Second, compared to plain net ResNet reduces the error percentage.

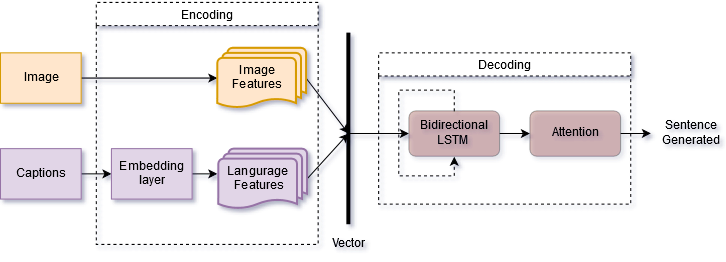
**CHAPTER - 3**

**PROPOSED SYSTEM**

**3.1 Architecture**

The architecture for any image captioning system consists of 3 main components. Figure 3.1 gives the proposed architecture.

* Encoding
  + Image features extraction
  + Caption sequence processing
* Decoding

****

**Figure 3.1 - Architecture**

First, we present remote sensing image representation methods (encoder). Then, the sentences generating models are introduced (decoder). The first method utilizes a deep multimodal neural network to generate a coherent sentence for a remote sensing image. We introduce the method in three parts: representing remote sensing image, representing sentences, and sentences’ generation.

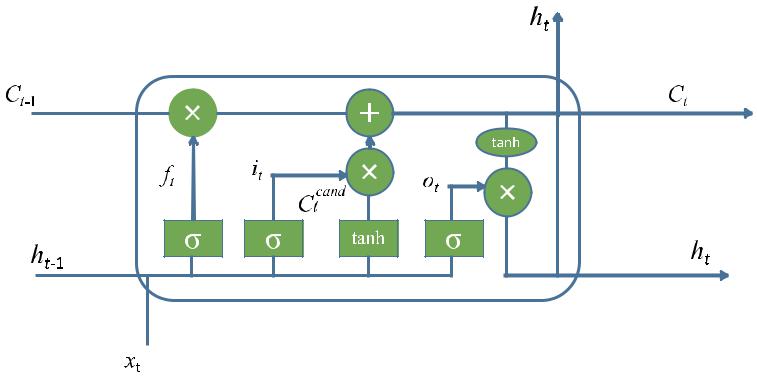
*Representing Remote Sensing Images*: The images are represented in the form of features required by the application. Extracting features helps to identify the minimum parts needed to process the application and avoiding unwanted detail. Various CNN algorithms are used to extract features from the dataset of given images. The deep features extracted from CNNs have found popular applications in Image Processing. The choice of algorithm is based on the specific application but newer algorithms are mostly preferred for efficiency. We use image pretrained features of CNNs – VGGNet, InceptionV3 , ResNet, pretrained on ImageNet along with Classification weights of the UCM Dataset. Inception uses sparse connections leading to reduction in computation load reducing bottleneck. ResNet or Residual Net helps to solve vanishing gradient by representing weights in the form of residue from previous layers. ResNet152 has the least error rate of 3.6%. The image representation can be given by

where *e* and *ed* represents the extracted features from images for multimodal and domain-driven multimodal methods respectively. *f*CNN and *f*c represents the function that extracts features from an image *i* and classification weights for the dataset *i*.

*Representing Sentences*: In the first method, every word in a sentence is represented by a one-hot K -dimensional word vector wi, where K is the size of the vocabulary. Then the word vector is projected to an embedding space by the embedding matrix Es. Es is an h × K matrix, where h is the dimension of embedding space. Sentence y is encoded as a sequence of h-dimensional projected word vectors.

*Sentences Generation*: In this section, a special RNN, called LSTM networks, is exploited to generate the sentences. This is because LSTM networks are more complicated than original RNNs. Sentence generating process is more like a “human,” which means that the human thoughts have persistence, and the next word can be predicted according to previous words to some extent. LSTM is designed to address the long-dependency problem using gates to control the information passed through the networks. The structure of LSTM is shown in figure 3.1.2. The first step of LSTM is using a forget gate to decide what information to throw away. Then, an input gate decides the values we will update, and a tanh layer is used to generate candidate values that could be added to the state of networks. The output is the cell state filtered by an output gate.

The process of sentence generating is considered as a problem that maximizes the correct sentence generating probability conditioned on the given image information.



**Figure 3.2 LSTM structure**

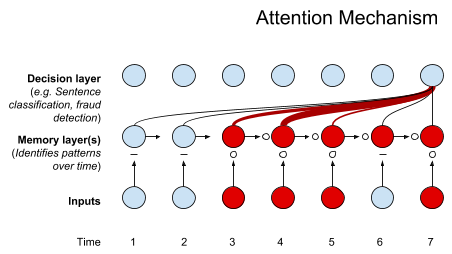
Initial research for image captioning consisted of constructing hand-crafted features based on image properties like their histogram, color model, etc.… for extracting features from images. As neural networks improved allowing faster computation rates and better results, the shift took place towards neural networks for processing images efficiently. Our experiment models consisted of following components.

**Table 3.1 Experiment Models**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Encoding** | | | **Decoding** |
| **Image Features extraction** | **Language Model** | **Class features** |
| *Baseline Model* | * VGGNet19 * ResNet50 * Inception V3 | Embedding Layer  LSTM (to encode) | None | Dense Layer |
| *LSTM Decoding* | * InceptionV3 * ResNet152 | Word Embedding  LSTM (to encode) | None | LSTM (to decode) |
| *Attention-based LSTM* | ResNet152 | Word Embedding  LSTM (to encode) | None | * Bidirectional LSTM * Attention Model |
| *Domain-driven attention based LSTM* | ResNet152 | Word Embedding | Added from classification using ResNet152 | * Bidirectional LSTM   Attention Model |

**3.2 Attention based LSTM**

Attention is another popular addition in an Image Captioning system. The decoding part in the Encoder-Decoder based models is used to convert the fixed length representations obtained from encoding into intermediate representations. The intermediate representations that give our expected output maybe of variable length. Attention as meaning suggests, allows decoder to identify where to look for, in sequence, from the given image. This helps to generate better coherent sentences. If the task is to generate caption for an image containing “a dog playing with a ball”, it is important to detect dog first and ball next, compared to all other irrelevant objects in the image. Attention mechanism provides ability for such preferential treatment with the help of probabilities, giving more weight to the objects dog and ball for that particular image. The addition of Attention mechanism coupled with ResNet152 for training images and using word embeddings for language model gave a result of BLEU-4 score of 0.610643.



**Figure 3.3 Attention Mechanism**

**3.3 Image Classification**

The primary task of image classification is acceptance of input image and the definition of its following class. Image classification is the problem of placing an image under a class based on what forms major part of the image. A class represents a tagline under which the given image can be put under. Image classification is a supervised learning problem, it defines a set of target classes and to train a model to recognize them using labeled example photos. CNN is the most recently preferred method for classification as it provides the advantage of automatic feature extraction. The image is passed through a series of convolution, non-linear, pooling and fully connected layers and then generates the output. The fully connected layer takes the output information from the convolution networks. Attaching the fully connected layer to end of the network results in an N dimensional vector, where N is the amount of classes from which the model chooses the desired class.

**3.4 Domain-Driven method**

The target classes are “domains” mentioned in our method. The domains refer to the different domains the images will fall upon classification. In our method, the fully connected layer is removed and the output of the previous layer which, for a given image, contains the probabilities of falling under different target classes present is used. The probabilities or weights are multiplied with our CNN extracted features to amplify the probability of finding the relevant word relating to the image for the generated sentence. This provides better results for our dataset since it consists of a lot of information helping in converging the important part of an image.

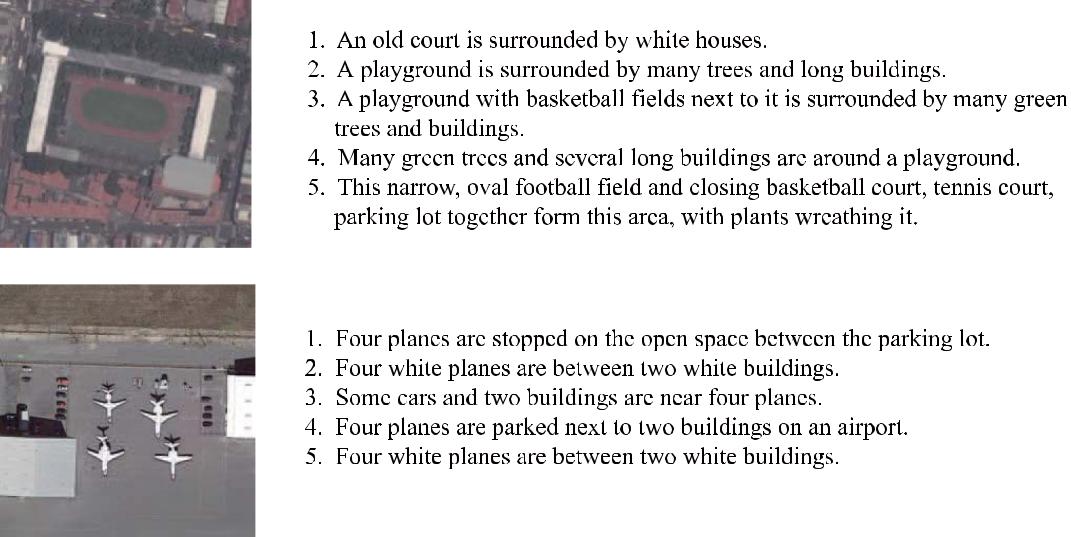
**CHAPTER – 4**

**IMPLEMENTATION DETAILS**

**4.1 Dataset**

**4.1.1 UCM-Captions Data Set**

This data set is based on the UC Merced Land Use data set. It contains 21 classes. The dataset images are land use image, including agricultural, airplane, baseball diamond, beach, buildings, chaparral, dense residential, forest, freeway, golf course, harbor, intersection, medium residential, mobile home park, overpass, parking lot, river, runway, sparse residential, storage tanks, and tennis court. For each class, there are 100 images. The shape of the images is 256 × 256 pixels. For each image, there are 5 captions to describe the image. The diversity of five coherent sentences for one image is totally different, but the sentence difference between images of the same class is very small. The size of the vocabulary is 293. Dividing the 2100 images into training, testing and validation sets, the number of image, caption pairs in each will be 1260, 210 and 210 respectively. The description length, which denotes maximum length of the sentences which in turn will be used to limit the maximum length of the sentence generated as output, is 24.



**Figure 4.1 – Dataset Example**

**4.1.2 UAVIC Dataset**

This data set is based on the drone images and videos collected from Team Dhaksha of MIT. The number of unprocessed images collected is 840 and unprocessed video is 2. The dimensions of the image are 6000 x 3376. The images were processed using OpenCV in python. The images were divided into 2 and useful portions were identified to be resized to the size of 400 x 400. The video was processed to identify useful information in the frames that can be used for dataset. The number of classes identified is 12. The identified dataset images are land images, including barren land, ground, farmlands, forests, garden, highways, playground, residential, road, runway, solar panel, water body and temple. For each class, there are 150 images. The shape of the images is 400 x 400 pixels. For each image, there are 5 captions to describe the image. The diversity of five coherent sentences for one image is totally different, but the sentence difference between images of the same class is very small.



**Figure 4.2 Unprocessed Image Figure 4.3 Processed Image**

**4.2 LSTM**

The model was refined by including LSTM for decoding vectors. It is an approach used in Image Captioning applications involving temporal dependencies. Using LSTMs makes it useful to generate coherent sentences. A sentence of the form “The cat chased the dog” has a totally different meaning than “The dog chased the cat”. The placement of a word in a sentence makes a big difference. In the baseline model, InceptionV3 provided better results than the others with a BLEU-4 score of 0.2248. With the addition of LSTM to the InceptionV3 trained model, the accuracy is 0.5032.

**4.3 Bidirectional LSTM**

Bidirectional LSTMs are an extension of traditional LSTMs that can improve model performance on sequence classification problems. Bidirectional LSTMs train two instead of one LSTMs on the input sequence. The first on the input sequence as it is and the second on a reversed copy of the input sequence. It learns the sentence dependence both forward and backward. This can provide additional context to the network and result in faster and even fuller learning on the problem. The attention mechanism is implemented with the help of bidirectional LSTM and the resulting BLEU-4 score is 0.610643.

**4.4 Word Embeddings (word2vec)**

Word Embeddings are the texts converted into numbers and there may be different numerical representations of the same text. Word2vec is one algorithm for learning a word embedding from a text corpus. There are two main training algorithms that can be used to learn the embedding from text; they are continuous bag of words (CBOW) and skip grams. The way CBOW work is that it tends to predict the probability of a word given a context. A context may be a single word or a group of words. Skip – gram follows the same topology as of CBOW. It just flips CBOW’s architecture on its head. The aim of skip-gram is to predict the context given a word. One advantage of word2vec is that words having similar meaning have closely similar values based on the degree of similarity. This will improve results in the language processing part.

**CHAPTER - 5**

**RESULTS AND DISCUSSION**

**5.1 Experimental Results**

The Image Captioning model obtained by various methodologies/architectures and trained using training images is evaluated using test set of images by generating captions using trained model and calculating accuracy. We evaluate multimodal method based on different kinds of features with randomly 80% for training, 10% for validation, and 10% for test on UCM-captions data set. We first test four handcrafted representations for captioning, and then use the different CNNs.

**5.1.1 BLEU score**

There are different metrics to evaluate the generated caption and BLEU score is one such metric. BLEU stands for Bilingual Evaluation Understudy. The BLEU score measures the co-occurrences of n-gram between generated and reference sentences. BLEU-1 is for checking 1-grams, BLEU-2 is for checking 2-grams and like till BLEU-4.

**5.2 Comparison with other results**

To evaluate the generated sentences based on handcrafted representations, four handcrafted representations are conducted including SIFT, BOW, FV, and VLAD. Each remote sensing image sized 224 × 224 is segmented evenly to 16 patches sized 56 × 56. For each patch, a SIFT feature is obtained by principal component analysis of the origin SIFT features. Finally, 16 SIFT features are concatenated into a vector to represent the image. Other handcrafted representations are based on SIFT. For BOW representation, the dictionary size is 1000. Below is the results of BLEU score obtained in methods involving handcrafted methods. For all four handcrafted representations, VLAD performs the best on UCM-captions data set.

**Table 5.1 Results of handcrafted methods**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  | **BLEU-1** | **BLEU-2** | **BLEU-3** | **BLEU-4** |
| **RNN[10]** | SIFT | 0.57303 | 0.44354 | 0.37963 | 0.33883 |
| BOW | 0.41067 | 0.22494 | 0.14519 | 0.10951 |
| FV | 0.5908 | 0.46026 | 0.39681 | 0.45446 |
| VLAD | 0.63108 | 0.51928 | 0.46057 | 0.42087 |
| **LSTM[11]** | SIFT | 0.55168 | 0.41656 | 0.34891 | 0.30403 |
| BOW | 0.39109 | 0.18767 | 0.10892 | 0.07058 |
| FV | 0.58972 | 0.46678 | 0.40799 | 0.36832 |
| VLAD | 0.70159 | 0.60854 | 0.54961 | 0.50302 |

The evaluation results for the base paper using Encoder-Decoder based architecture is

**Table 5.2 Results on Base Paper**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **BLEU – 1** | **BLEU – 2** | **BLEU – 3** | **BLEU - 4** |
| VGG16 | 0.74569 | 0.65976 | 0.59485 | 0.53932 |
| InceptionV3 | 0.83751 | 0.76217 | 0.70420 | 0.65624 |
| AlexNet | 0.78498 | 0.70929 | 0.65182 | 0.60167 |

The result shows that all the CNNs features are better than handcrafted features. All the CNNs features get almost the same result. In CNNs features, AlexNet gets the best result on BLEU. The results consolidated from the reference papers is

**Table 5.3 Results on Reference Paper**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Method** | **Dataset** | **BLEU-1** |
| 1 | RNN[4] | Flickr8K | 0.579 |
| 2 | Show & Tell Model[5] | MSCOCO | 0.655 |
| 3 | Convolutional approach[6] | MSCOCO | 0.725 |
| 4 | CNN+CNN[7] | Flick30k | 0.607 |

The results are compared with the evaluation results of the baseline model implementation. The results for different changes in architecture are consolidated as follows.

**Table 5.4 Comparison of results with the baseline model**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **BLEU – 1** | **BLEU – 2** | **BLEU – 3** | **BLEU - 4** |
| ResNet152 | 0.74569 | 0.65976 | 0.59485 | 0.53932 |
| InceptionV3 with LSTM | 0.762451 | 0.669016 | 0.632124 | 0.543431 |
| ResNet152 with LSTM | 0.80078 | 0.708322 | 0.668718 | 0.581832 |
| ResNet152 with Attention- based Bi-directional LSTM | 0.812111 | 0.729575 | 0.693783 | 0.610643 |

The following observations are made from the above table. Adding a LSTM layer improves the accuracy of the model. Using ResNet152 for features extraction provides better result than the others (VGG19, ResNet50, and InceptionV3). Similarly applying Bi-directional LSTM along with Attention improves the accuracy and provides us with coherent sentences. This is because we know Attention helps to prioritize the objects to look for.

**5.3 Ground Truth Comparison**

To check if the generated caption is comparable and able to describe the image, it is compared with ground truths as show in the figures. The sentences generated where closely similar to the ground truths as in case of Resnet152 with Attention based Bi-directional LSTM.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **BLEU – 1** | **BLEU – 2** | **BLEU – 3** | **BLEU - 4** |
| UCM Dataset | 0.812111 | 0.729575 | 0.693783 | 0.610643 |
| UAVIC Dataset | 0.649809 | 0.41589 | 0.296616 | 0.170059 |
| Combined | 0.705707 | 0.570448 | 0.519331 | 0.421084 |

**Table 5.5 Comparison of results from different combinations of dataset model**

The results present in the above table is that of the model when trained using only UCM Dataset , only UAVID Dataset and both combined. The results are based on evaluating captions from both the datasets. The result of classification is

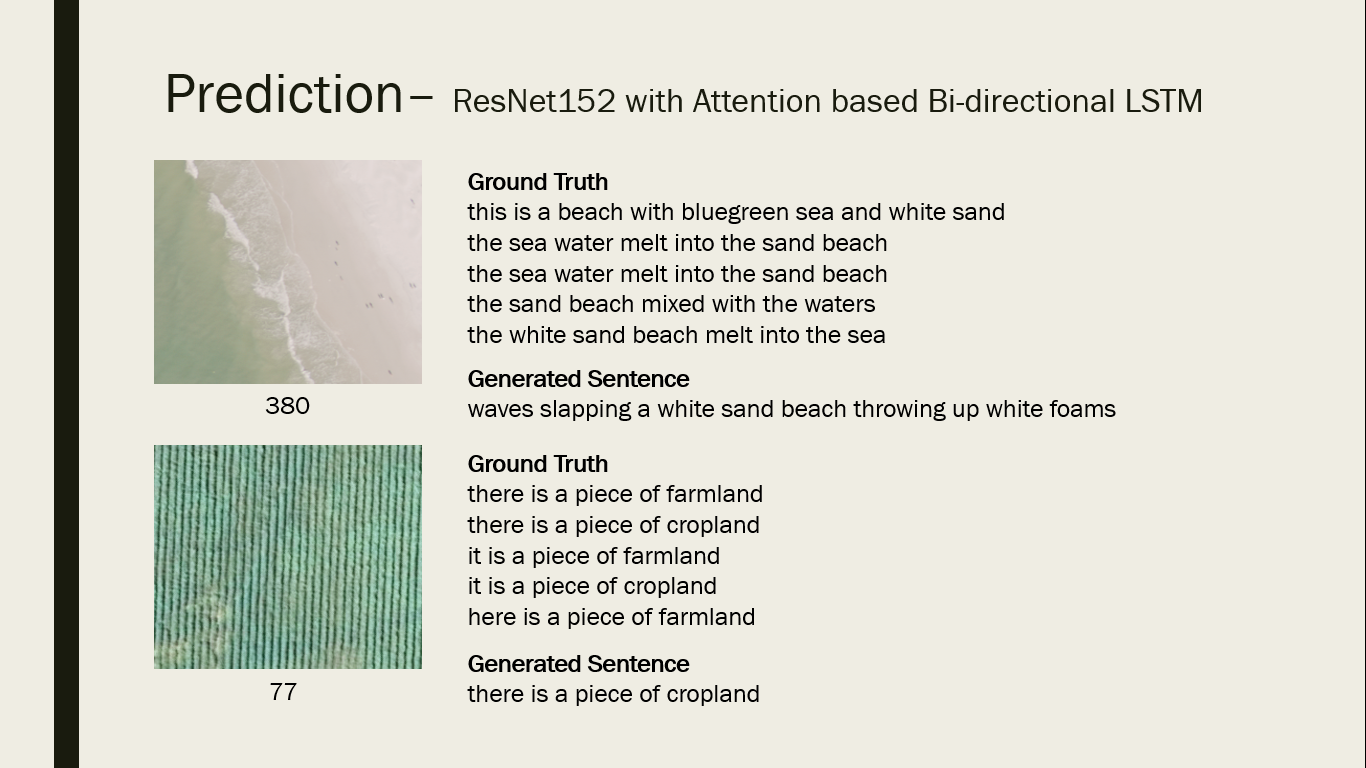
|  |  |  |
| --- | --- | --- |
| **Model** | **Train Accuracy** | **Test Accuracy** |
| UCM Dataset | 1.00 | 0.85 |
| UAVIC Dataset | 0.25 | 0.21 |
| Combined | 0.68 | 0.55 |

The classification model weights are applied as an additional input to the model which as a result of multiplication, increases the class probability which in turn helps in producing better captions. The results are improved as follows,

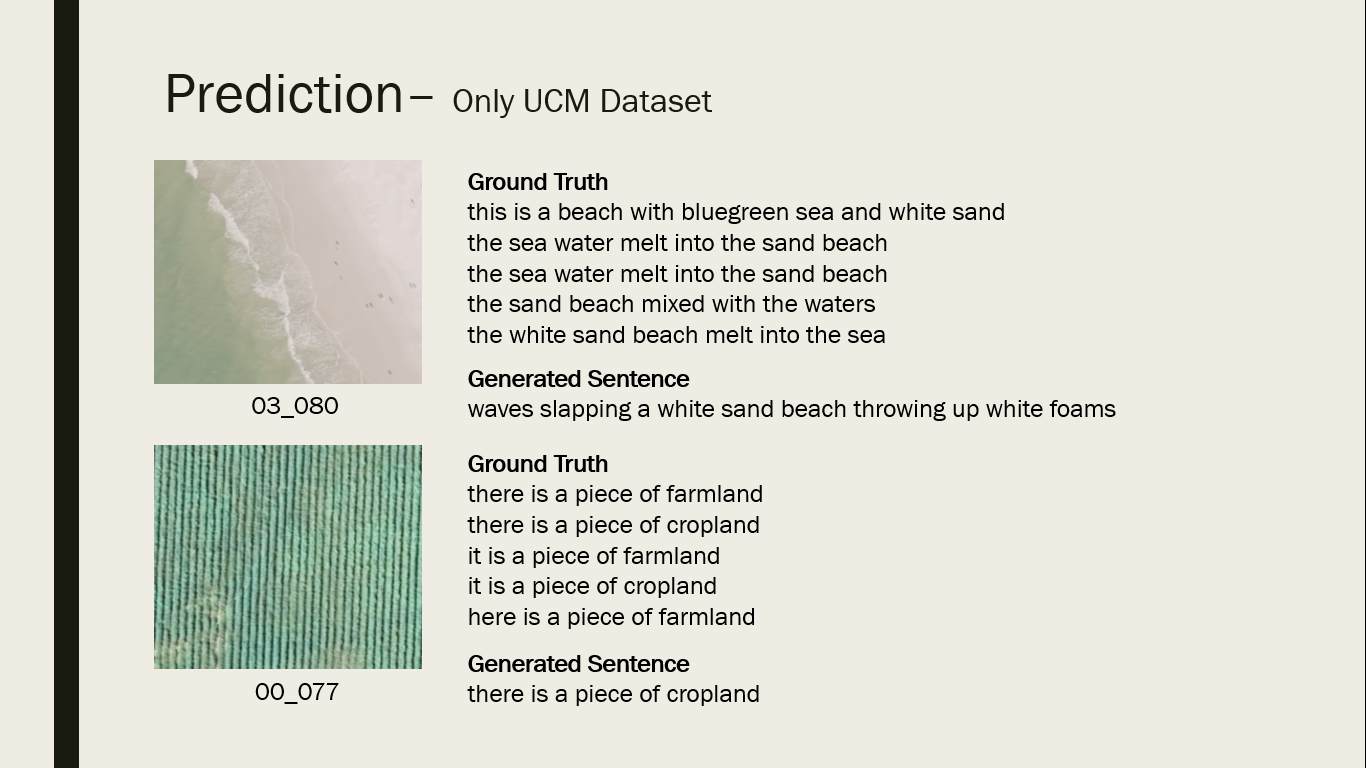
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **BLEU – 1** | **BLEU – 2** | **BLEU – 3** | **BLEU - 4** |
| UCM Dataset | 0.841121 | 0.745565 | 0.683783 | 0.620143 |
| New Dataset | 0.749809 | 0.51589 | 0.396456 | 0.290019 |
| Combined | 0.755707 | 0.590448 | 0.559345 | 0.481025 |

**Table 5.6 Comparison of results from different combinations of dataset model**

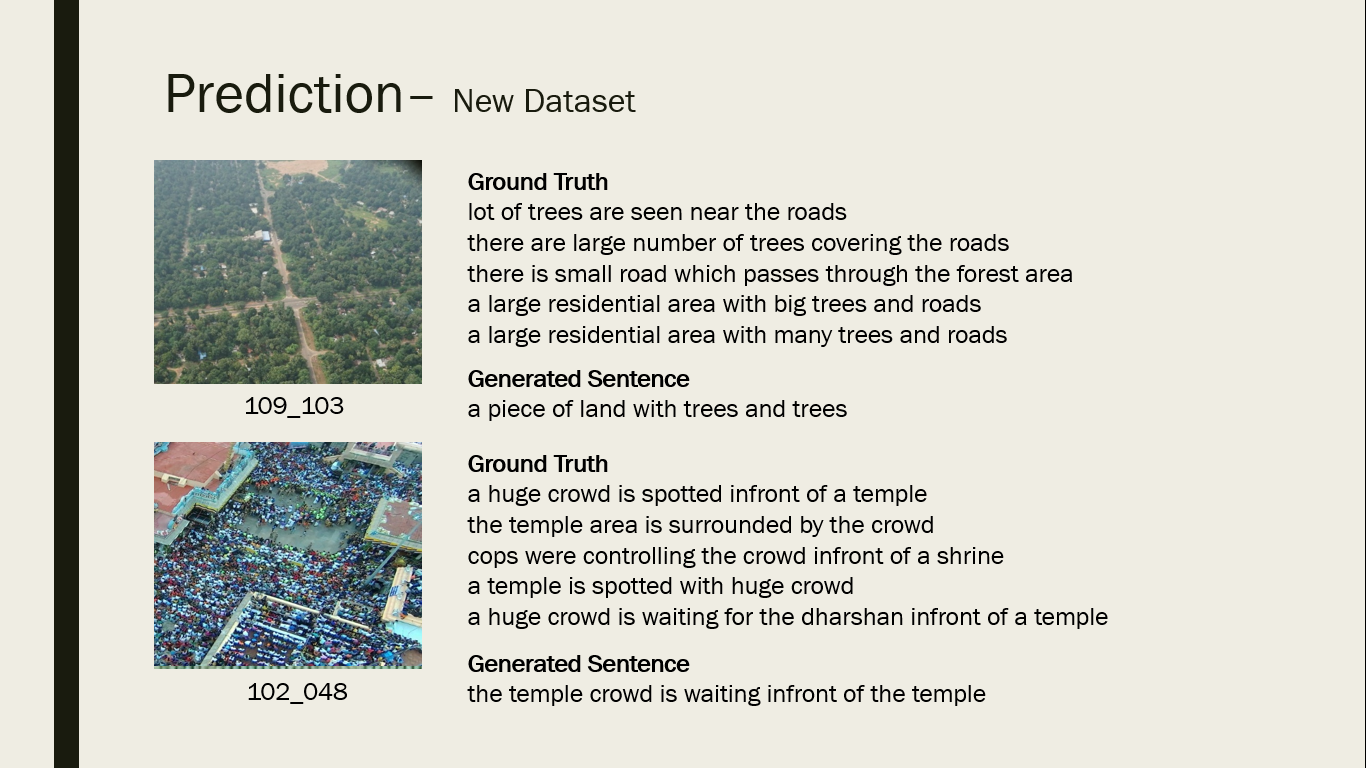
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Model** | | **BLEU – 1** | **BLEU – 2** | **BLEU – 3** | **BLEU - 4** |
| **Train** | **Test** |
| UCM Dataset | UCM | 0.841121 | 0.745565 | 0.683783 | 0.610143 |
| UAVIC |  |  |  |  |
| Combined |  |  |  |  |
| UAVIC Dataset | UCM | 0.3249860 | 0.104175249 | 0.043349196 | 0.06006527 |
| UAVIC | 0.6360332 | 0.403450039 | 0.294971207 | 0.181209161 |
| Combined | 0.485145041 | 0.268615919 | 0.188338416 | 0.101146428 |
| Combined | UCM | 0.755707 | 0.590448 | 0.559345 | 0.481025 |
| UAVIC | 0.56850152 | 0.337477049 | 0.243249131 | 0.135271849 |
| Combined | 0.673459659 | 0.523108281 | 0.460512249 | 0.351685143 |

­­

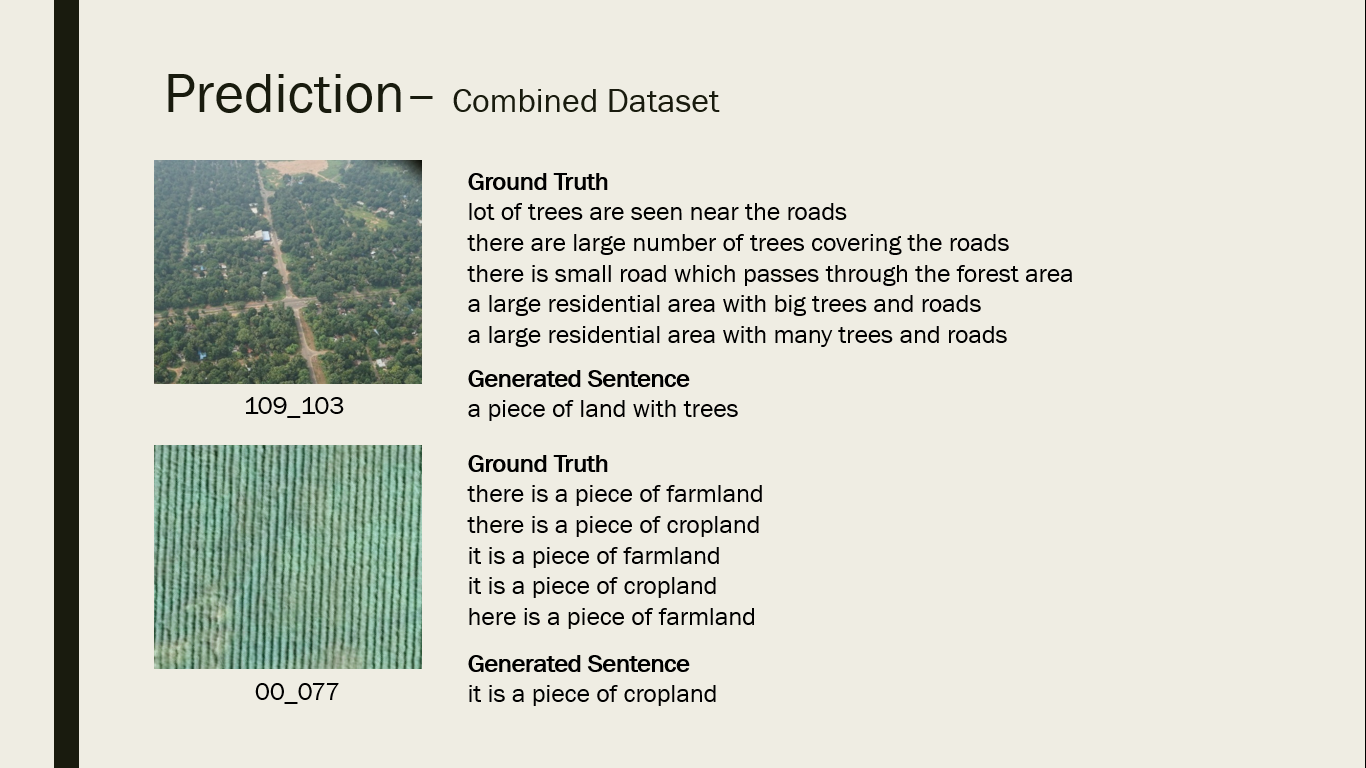
**Figure 5.1- ResNet152 with attention based B-LSTM**



**Figure 5.2 – UCM Dataset**



**Figure 5.3 – New Dataset**



**Figure 5.4 – Combined Dataset**

**CHAPTER - 6**

**CONCLUSION AND FUTURE WORK**

**6.1 Conclusion and Future Work**

We propose a novel deep learning approach to the problem of remote sensing image captioning. Our model successfully extracts features from the image and generates sentences based on convolutional features and it is found that the model performs the best for ResNet 152 with BLEU-4 accuracy of 61.06%.The experimental results show that the image caption methods for natural image can be transferred to remote sensing image to obtain only acceptable descriptions. But considering the characteristics of remote sensing images, more works need to be done on remote sensing image caption generation. In future work, the dataset will be more comprehensive than the present version because some of the sentences are obtained by duplicating the existing sentences. And we plan to apply some new techniques in image processing field and natural language processing field to remote sensing image caption generation task.

**CHAPTER-7**

**REFERENCES**

[1] V. Ordonez, G. Kulkarni, and T. L. Berg, “Im2text: Describing imagesusing 1 million captioned photographs,” in Proc. Adv. Neural Inf. Process. Syst., 2011, pp. 1143–1151.

[2] S. Li, G. Kulkarni, T. L. Berg, A. C. Berg, and Y. Choi, “Composing simple image descriptions using Web-scale n-grams,” in Proc. 15th Conf. Comput. Natural Lang. Learn., 2011, pp. 220–228.

[3] Y. Yang, C. L. Teo, D. H. Iii, and Y. Aloimonos, “Corpus-guided sentence generation of natural images,” in Proc. Conf. Empirical Methods Natural Lang. Process., 2011, pp. 444–454.

[4] A.Karpathy and L.FeiFei, “Deep visual-semantic alignments for generating image description” in Proc.IEEE conference Computer Vision Pattern Recognition, June 2015.

[5] Parth Shah, VishvajitBakarola and SupriyaPatti,“Image Captioning using Deep Neural Architectures”,January 2018.

[6] Jyoti Aneja, Aditya Deshpande, Alexander G.Schwing, ”Convolutional Image Captioning” .

[7] QingZhong wang and Antoni B.chan, ”CNN+CNN : Convolutional Decoders for Image Captioning,” May 2018.

[8] Z. Shi and Z. Zou, “Can a machine generate humanlike languagedescriptions for a remote sensing image?” IEEE Trans. Geosci. RemoteSens., vol. 55, no. 6, pp. 3623–3634, Jun. 2017.

[9] Justin Johnson, Andrej Karpathy and Li Fei-Fei, “DenseCap:Fully Convolutional Localization Networks for Dense Captioning,” arXiv:1511.00757, 24 Nov 2015.

[10] R. J. Williams and D. Zipser “A learning algorithm for continually running fully recurrent neural network,” Neural comput., vol. 1, no. 2,pp. 270-280, 1989.

[11] S.Hochreiter and J.Schmidhuber, “Long short-term memory,” Neural comput., vol. 9, no. 8,pp. 1735-1780, 1997.

[12] Kelvin Xu, Jimmy Lei Ba, Ryan Kiros and Richard S. Zemel, “Show, Attend

and Tell: Neural Image Caption Generation with Visual Attention,”

arXiv:1502.03044, 19 Apr 2016.

[13] Moses Soh, “Learning CNN-LSTM Architectures for Image Caption Generation”.

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren and Jian Sun, “Deep Residual Learning for Image Recognition”. arXiv:1515.03385, 10 Dec 2015.