Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 165 (2019) 32–40

Procedia Computer Science

www.elsevier.com/locate/procedia

INTERNATIONAL CONFERENCE ON RECENT TRENDS IN ADVANCED COMPUTING 2019, ICRTAC 2019

Region Driven Remote Sensing Image Captioning

S Chandeesh Kumar*, Hemalatha M, S Badri Narayan, P Nandhini

Anna University, Chennai, Tamil Nadu, India

Abstract

Remote sensing Image Captioning is a special case of Image Captioning which solves the difficulties in processing the remote sensing images. Issues may arise due to translation, rotation and viewpoint of images and maintaining semantic consistency in the generated captions. This method of describing a remote sensing scene in the form of sentences plays an important role in a number of fields, such as image retrieval, scene classification and as a vision companion. A Domain-driven approach is developed, in which the domain probabilities are used for captioning the remote sensing images. This approach concentrates on the domain-based information available in the images. A new dataset, called UAVIC dataset is created for images captured using Unmanned Aerial Vehicle (UAV), which covers wide range of land having multiple terrains and gives a better view of the landscapes. The proposed domain driven approach is applied to UCM and UAVIC dataset and the quality of resulting captions are evaluated using BLEU scores.

Keywords: UAV; image processing; remote sensing; caption generation

1. Introduction

Image captioning is the method of generating image descriptions 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. 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 images has their application in the assessing the quality of a land - if it is fertile, infertile or barren,

* Corresponding author. Tel.: 8754564990
E-mail address: taekchan331@gmail.com

10.1016/j.procs.2020.01.067

weather conditions, traffic/crowd detection, classifying the type of activity taking place at a location- if it is an illicit activity or allowable activity. The challenges in captioning a remote sensing image are as follows.

- *Scale Ambiguity*: The objects in the ground have different scales under different views. (Mostly top view)
- *Category Ambiguity*: Objects maybe fused together and hence may fall under different categories.
- *Rotation Ambiguity*: The remote sensing image considered can be in different rotated angle or translation. There is no up, down, left and right in a remote sensing image.

Remote Sensing image captioning is a specialized application narrowed by the type of images used. The output can be described in the formed of well-developed sentences to give crisp details about land for 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 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. The existing dataset is created in such a way that it is computer friendly. It contains mostly single objects present in a easily identifiable way. The new dataset containing images captured by drones resemble close to real time, is human friendly and contains diverse variety of objects. This means the model has to be more efficient to find out the right object and right word to summarize the image. 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 in order to be processed by the models. There are different methods for converting words to integers and processing the words through the integers. In general, the image captioning problem is the problem of generating a word, given picture and previous word. This can be achieved with the help of RNNs, LSTM and text processing libraries like word2vec. The organization of the paper is as follows - Section 2 discusses the literature survey. Section 3 the proposed framework. Section 4 describes the implementation details. Section 5 discusses evaluation of the region driven approach. Section 6 presents the conclusion.

2. Literature Survey

2.1. Natural Image Captioning

2.1.1. Handcrafted Features

The handcrafted feature methods use handcrafted methods to extract features from the images. This involves using different encoding techniques to obtain an image representation. In [1] retrieval-based method is proposed that first retrieves images related to a query image. The caption of the retrieved images which are similar to the query image are used to generate captions. As the retrieved images are used in generating captions, if the images are retrieved incorrectly then the captions become irrelevant. The captions generated using the retrieval based method are grammatically incorrect, which makes the captions hard to understand. The method performs poorly when the given image is sparsely closer to the images in the dataset.

The object detection based approaches are proposed in [2], [3]. In these types of methods, the various objects in the image is identified first and then the association between them is modelled. The caption is generated using a text model which generates captions based on the association between objects. It is found that this approach a very good object detection model to generate a better caption. It is also seen that the quality of the finally generated caption is largely dependent on the quality of the object detection.

2.1.2. Neural Network based Methods

A neural network based model is proposed in [5] a model that 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 is described 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. Limitations in this model is that it can only generate description of only one input array of pixels at a fixed resolution.

[6] proposed a model that automatically creates description for images using natural languages. The proposed approach uses Show & Tell model. The model [6] uses the combination of Inception-v3 and Long Short-Term Memory (LSTM) cell and Inception-v3 provides object recognition capability. The show and tell model uses the capacity of CNN and LSTM to generate captions for the images.

Aneja et al. developed a convolutional image captioning technique in [8]. The work analyzed the CNN and RNN approaches and concluded that CNN produces more entropy in the output probability distribution. The drawbacks in RNN-based techniques are, the training process are sequential for a particular image-caption pair. Secondly RNN produces lower classification accuracy and still suffer from vanishing gradient problem.

2.2. Attention

An attention-based model is developed that automatically learns to describe the content of images. In addition to CNN, attention mechanism was introduced to attend to region of interest. The approach used in [17] computed separate attention parameters and a separate attended vector for every word. The work involved performing beam search with different beam sizes and found that for both LSTM and CNN methods the performance increases.

The approach proposed in [7] developed a CNN+CNN framework for image captioning. [7] proposed a framework that only employs convolutional neural networks (CNNs) to generate captions. 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.

The approach in [10] performs dense captioning for images. The method proposed developed an identifies the regions of interest of the given images and the describes each region with natural language. [10] proposed 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. The localization layer identifies spatial regions of interest and smoothly extracts a fixed-sized representation from each region. The proposed model is an encoder-decoder approach combined with attention mechanism. For encoding part [13] used a convolutional neural network (CNN) to extract a feature vector. For decoding, they used Long Short-Term Memory (LSTM). The LSTM [14] generates a single word at each time step given the previous hidden state and also the words generated previously. Attention tells where exactly to look when the neural network is trying to predict parts of a sequence. The focusing region is done by random sampling. However, the expected region may have higher chance to be focused on than the others.

2.3. Remote Sensing Image Captioning

In this work we propose to use image captioning on Remote Sensing images. There are only very few approaches for captioning the remote sensing images. Shi et al. in [9] tried to describe the remote sensing images like a human. Humans describes an image in levels of key objects, environment and landscape. The approach proposed in [9] 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.

3. Region Driven Captioning Framework

The framework of the proposed Region Driven Remote Sensing Image Captioning approach is shown in Figure 1. The proposed architecture consists of 2 modules - Encoding and Decoding parts. In encoding, features are extracted from images and caption sequence is processed using text processing. In Decoding, the encoded features are decoded to generate captions for given images.

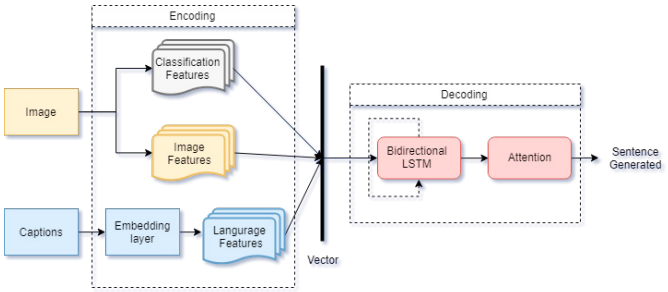


Fig. 1: Framework of the proposed Region Driven Remote Sensing Image captioning approach

3.1. Encoding

Image features extraction. The images are represented in the form of features required by the application. Various CNN algorithms are used to extract features from the dataset of given images. We use features extracted from pre-trained 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. ResNet 152 has the least error rate of 3.6%. The image representation can be given in equation(1)

$$e = f_{CNN}(I) \quad and \quad e_d = f_{CNN}(I) + f_c(I) \tag{1}$$

where e and e_d represents the extracted features from images for multimodal and domain-driven multimodal methods respectively. f_{CNN} represents the function that extracts feature from an image I and f_c represents classification weights for the dataset I.

3.2. Decoding

Long Short-Term Memory(LSTM). RNN is the first algorithm that remembers its input due to internal memory. Because of the presence of internal memory, it can be able to recollect significant details related to the input they received, which enables them to precisely predict the next term. RNN has short-term memory. In this work we use LSTM networks for sentence generation since it solves the long dependency problems, hence it controls the information passed from one state to another state. The sentence generating process generates the probability of current word given the previously generated words. Hence it can be considered as a mathematical problem of finding the conditional proba- bility. Figure 2 shows the structure of LSTM.

$$\begin{aligned} &f_t = \sigma(W_f \cdot [h_{(t-1)}, x_t] + b_f), \quad i_t = \sigma(W_i \cdot [h_{(t-1)}, x_t] + b_i), \quad o_t = \sigma(W_i \cdot [h_{(t-1)}, x_t] + b_o) \\ &C_t = f_t * C_{(t-1)} + i_t * C_t^{cand} \quad where\ C_t^{cand} = tanh(W_c \cdot [h_{(t-1)}, x_t] + b_c) \\ &h_t = o_t * tanh(C_t) \end{aligned} \tag{2}$$

In equation (2) f_t denotes the forget gate, i_t the input gate, o_t the output gate. The equations indicate the operations carried out in a LSTM layer for each operation. Each gate performs following functions: The input gate decides the value to be updated, it is represented by i_t. The *tanh* layer generates the values to be added to the state of the cell, it is represented by C_t^{cand}. The output gate filters the cell state and produces the output, denoted by o_t.

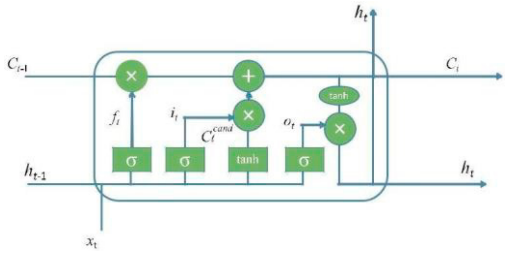


Fig. 2: Structure of LSTM [16]

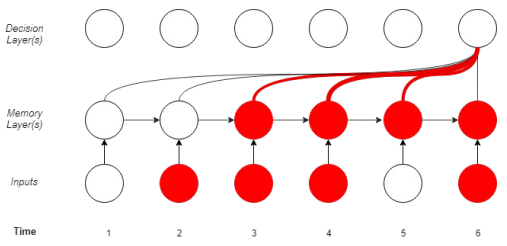


Fig. 3: Attention Mechanism

3.3. 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 mechanism allows decoder to identify where to look in from the given image, which helps to generate better coherent sentences. Attention mechanism provides ability for such preferential treatment with the help of probabilities, giving more weight to the objects for a particular image. Attention mechanism is shown in the Figure 3.

Bidirectional LSTMs are an extension of traditional LSTMs that can improves the performance of the model on sequence classification problems. Bidirectional LSTMs train two LSTMs. 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 on the output of the bidirectional LSTM.

3.4. Region based image classification

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 grouped. CNN is the most recently preferred method for classification as it provides the advantage of automatic feature extraction. The given image is passed through a number of convolution, non-linear, pooling and fully connected layers and then the output is generated at the final layer. The output information from the previous convolution layers in the neural networks is used by the fully connected layer. The purpose of adding a fully connected layer at end of the convolutional neural network is to obtain an N dimensional vector, where N is number of classes from which the model outputs the desired class.

In this method, domains are the classification classes to which images belong to. Some of the domains are Forests, Buildings, Temples, etc. 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 guides the model in type of the landscape present in the image.

4. Experiments

4.1. Dataset

UCM Dataset. This dataset is developed using the data from UC Merced Land Use given in [16]. It consists of 21 classes including land use image, including agricultural, etc. 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. It is found that the captions of images belonging to the same class is similar. The 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 givesan example of the UCM dataset containing the image along with its ground truths in the form of 5 captions.

1. An old court is surrounded by white houses.
2. A playground is surrounded by many trees and long buildings.
3. A playground with basketball fields next to it is surrounded by many green trees and buildings.
4. Many green trees and several long buildings are around a playground.
5. This narrow, oval football field and closing basketball court, tennis court, parking lot together form this area, with plants wreathing it.

Fig. 4: UCM Datasets

The aerial view of forests alongside roads.
The growth of trees are more dense in that area.
There is a group of trees scattered.
There are many cell towers located in between the forests.
The area consists of green lands with some trees grown in it.

Fig. 5: UAVIC Dataset

Table 1: Number of images in UAVIC dataset

	Image Class	Total	Training	Testing
1	Barren lands	150	90	30
2	Farmlands	150	90	30
3	Forests	97	58	19
4	Gardens	52	31	10
5	Highways	150	90	30
6	Playgrounds	50	30	10
7	Residential	150	90	30
8	Roads	150	90	30
9	Runway	60	36	12
10	Solar Panels	150	90	30
11	Water Bodies	100	60	20
12	Temple	26	15	5

UAVIC Dataset. This data set is based on the drone images and videos collected from Team Dhaksha [15] 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 unprocessed image is processed and then categorized into several classes and then captions are developed. The processed drone images are of same high quality and their size have been reduced drastically. The size of the vocabulary of the UAVIC dataset is 940 compared to 293 in UCM dataset. This means that the captions contains a diverse variety of words in case of UAVIC dataset. The video was processed to identify useful information in the frames that can be used for dataset. The number of classes identified is 12. Table 1 shows the details of UAVIC dataset.

The shape of the images is 400 x 400 pixels. For each image, there are 5 captions to describe the image. These five captions are developed in such a way that it covers all the features in the given image. UAVIC dataset example is shown in the Figure 5. The captions are given in detail and with precise predictions. There is plethora of things which can be spotted in a particular image and five captions are developed in such a way that it covers all the objects and relationship between them in detail. Based on classes the captions are given to the approximation on how the user grasps the image and constructs the sentence for image relating the objects.

4.2. Implementation details

The environment used to implement the proposed approach is Keras in anaconda with TensorFlow as backend. For running resource intensive models, cloud platform Google Colab is used. We evaluate the proposed multi-modal method for features extracted using different methods. The dataset is randomly divided into 80% for training, 10% for validation, and 10% for test on UCM-captions data set and UAVIC dataset.

For text processing pre-trained word2vec [4] model is used. Word Embedding's is the intermediate text representation in the form of matrix of numbers and it is seen that the numbers present in the matrix need not be same for the same text. The word2vec generates word features such that the words having similar meaning have closely similar values based on the degree of similarity.

5. Results and Discussion

5.1. Experimental Results

The Image Captioning model is implemented by using various methodologies and architectures. The performance of the trained models is calculated using BLEU scores. We first evaluate four handcrafted representations for image captioning, and then try the different CNNs. Table 2 shows the different experimental models created by our work.

Table 2: Various Experiment Models with their approaches

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

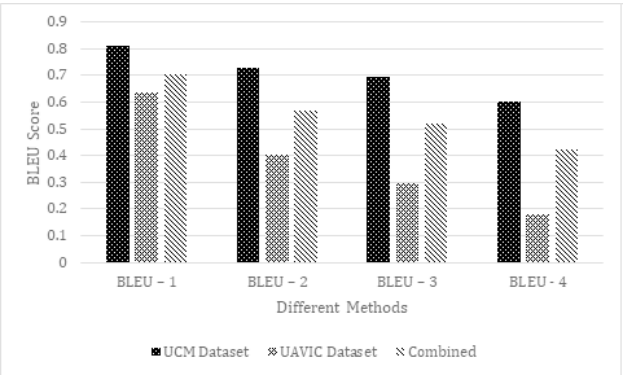


Fig. 6: Comparison with different combination of datasets - Domain-driven method

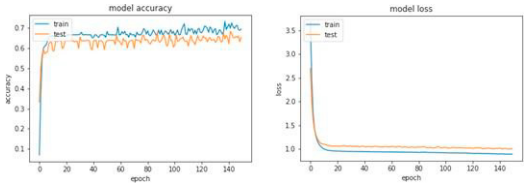


Fig. 7: Accuracy and Loss

Table 3: Classification Accuracy of different datasets

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

The generated caption is compared with the ground truth and it is inferred that the sentences are mostly similar to the ground truth in the case of the UCM dataset. The graphical representation of the results shown in Figure 6 is for the model trained using only UCM Dataset, only UAVIC Dataset and both combined. The results are based on evaluating captions from both the datasets. The Figure 7 shows the training losses and accuracy along the progression of each epoch on applying the proposed approach to combined datset. The classification model weights are applied as an additional input to the model which as a result of multiplication. It increases the class probability which in turn helps in producing better captions. Table 4 provides the results obtained for comparison of different combination of datasets. Table 7 describes the classification results of different datasets. In UCM dataset, the images are cropped with manually defined borders in such a way that it is easy to identify objects. UAVIC consists close to real drone images containing overlapping classes and more ambiguities. It is seen that the highest BLEU score better for UCM dataset. On observing the generated sentences and comparing it with images it is found that the generated sentences is more informative in case of training using Combined and testing using UCM dataset. So the caption generated using both the datasets contains information from both the UCM and UAVIC datasets. Since UAVIC dataset contain more like a real-time image, the caption is close to real-time caption. Also, the UAVIC dataset captions are detailed and specific about the attributes present in the image. The following observations are made from the Table 5, which compares different experimental models.

- 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).
- 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.

Table 4: Comparison of results on different dataset

Model		B-1	B-2	B-3	B-4
Train	*Test*				
UCM	UCM	0.84	0.74	0.68	0.61
	UAVIC	0.48	0.19	0.10	0.04
	Combined	0.58	0.38	0.32	0.22
UAVIC	UCM	0.32	0.10	0.04	0.06
	UAVIC	0.63	0.40	0.29	0.18
	Combined	0.48	0.26	0.18	0.10
Combined	UCM	0.75	0.59	0.55	0.48
	UAVIC	0.56	0.33	0.24	0.13
	Combined	0.70	0.57	0.51	0.42

Table 5: Comparison of results obtained using different Experimental models

Model	BLEU 1	BLEU - 4
ResNet152 [18]	0.74	0.53
InceptionV3 + LSTM	0.76	0.54
ResNet152 + LSTM	0.80	0.58
ResNet152 + Att- based Bi-LSTM	0.81	0.60
Region driven approach + ResNet152 + Att-based Bi-LSTM	**0.84**	**0.61**

5.2. Comparison with other results

To evaluate the generated captions based on handcrafted representations of image, four handcrafted representations are considered including SIFT, BOW, FV, and VLAD. The results of BLEU score obtained in methods involving handcrafted methods is given in Table 6 as given in [16]. From Table 7 it is seen that the caption model trained using CNNs features produce better scores than the caption models trained using handcrafted methods. It is shown that the proposed approach performs better than the other approaches to Remote sensing image captioning.

Table 6: Comparison of Handcrafted features-based methods on UCM dataset

	Methods	BLEU-1	BLEU-2	BLEU-3	BLEU-4
RNN [11]	SIFT	0.57	0.44	0.37	0.33
	BOW	0.41	0.22	0.14	0.10
	FV	0.59	0.46	0.39	0.45
	VLAD	0.63	0.51	0.46	0.42
LSTM [12]	SIFT	0.55	0.41	0.34	0.30
	BOW	0.39	0.18	0.10	0.07
	FV	0.58	0.46	0.40	0.36
	VLAD	0.70	0.60	0.54	0.50

Table 7: Comparison of results of Region driven approach with other approaches on UCM dataset

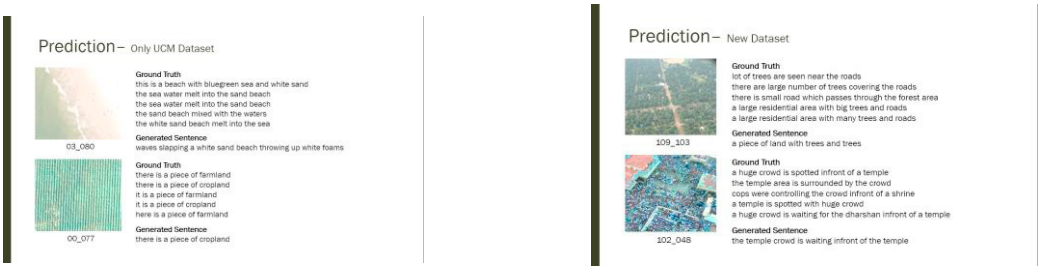
Model	BLEU 1	BLEU 3	BLEU - 4
VGG16 [20]	0.74	0.59	0.53
AlexNet [19]	0.78	0.65	0.60
Region driven approach	**0.84**	**0.68**	**0.61**

5.3. Sample Images and Captions Generated

Sample images and their descriptions are shown in this section. Figure 8a shows the captions generated using Region driven Remote sensing captioning on UCM dataset. Figure 8b shows the captions generated using Region driven Remote sensing captioning on UAVIC dataset. It is seen that most of the captions is approximately equivalent to the ground truth captions.

6. Conclusion

We proposed a novel deep learning method to generate captions for remote sensing images. Our method uses a region-driven approach for sentence generation. The proposed method efficiently extracts features from the image and generates captions based on convolutional features. The domain driven approach uses domain probabilities to caption a remote sensing image. The probabilities are multiplied with CNN extracted image features in order to improve the features in classification. It is found that the model performs the best for region-driven pretrained ResNet152 with BLEU-4 score of 61.06%. Hence it is inferred that the proposed approach helps to converge and increase the



(a) Sample results on UCM dataset (b) Sample results on UAVIC dataset

Fig. 8: Sample results obtained using proposed approach

importance of classification to identify the domain to which the image belongs. From the experiments on the remote sensing images, it is observed that more works is essential in the domain of remote sensing image caption generation. The dataset can be made more comprehensive than the present version by making the images focused on a single entity than providing it as a complete overview.

References

[1] Ordonez, Vicente and Kulkarni, Girish and Berg, Tamara L. (2011) "Im2text: Describing images using 1 million captioned photographs" *Advances in neural information processing systems* 1143–1151.
[2] Li, Siming and Kulkarni, Girish and Berg, Tamara L and Berg, Alexander C and Choi, Yejin. (2011) "Composing simple image descriptions using web-scale n-grams." *Proceedings of the Fifteenth Conference on Computational Natural Language Learning* 220–228.
[3] Yang, Yezhou and Teo, Ching Lik and Daume´ III, Hal and Aloimonos, Yiannis. (2011) "Corpus-guided sentence generation of natural images." *Proceedings of the Conference on Empirical Methods in Natural Language Processing* 444–454.
[4] Tomas Mikolov and Kai Chen and Greg S. Corrado and Jeffrey Dean (2013) "Efficient Estimation of Word Representations in Vector Space", *arXiv:1301.3781*
[5] Karpathy, Andrej and Fei-Fei, Li. (2015). "Deep visual-semantic alignments for generating image descriptions", *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3128–3137.
[6] Shah, Parth and Bakrola, Vishvajit and Pati, Supriya (2017) "Image captioning using deep neural architectures", *International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS)*, 1–4.
[7] Wang, Qingzhong and Chan, Antoni B (2018) "CNN+ CNN: Convolutional decoders for image captioning", *arXiv preprint arXiv:1805.09019*.
[8] Aneja, Jyoti and Deshpande, Aditya and Schwing, Alexander G (2018) "Convolutional image captioning", *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5561–5570.
[9] Shi, Zhenwei and Zou, Zhengxia (2017) "Can a machine generate humanlike language descriptions for a remote sensing image?", *IEEE Transactions on Geoscience and Remote Sensing*,**55** (6): 3623–3634.
[10] Johnson, Justin and Karpathy, Andrej and Fei-Fei, Li (2016) "Densecap: Fully convolutional localization networks for dense captioning", *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4565–4574.
[11] Williams, Ronald J. and Zipser, David (2015) "A Learning Algorithm for Continually Running Fully Recurrent Neural Networks", *Neural Computing, MIT Press*,**1** (2): 270–280.
[12] Hochreiter, Sepp and Schmidhuber, Jürgen (1997) "Long short-term memory", *Neural computation, MIT Press*, **8** (8): 1735–1780.
[13] Xu, Kelvin and Ba, Jimmy and Kiros, Ryan and Cho, Kyunghyun and Courville, Aaron and Salakhudinov, Ruslan and Zemel, Rich and Bengio, Yoshua (2015) "Show, attend and tell: Neural image caption generation with visual attention", *International conference on machine learning*, 2048–2057.
[14] Soh, Moses (2016) "Learning CNN-LSTM architectures for image caption generation", *Dept. Comput. Sci., Stanford Univ., Stanford, USA*.
[15] Dhaksha Team (2018) "Drone Manufacture in INDIA - Team Dhaksha", *https://www.teamdhaksha.com/*.
[16] Lu, Xiaoqiang and Wang, Binqiang and Zheng, Xiangtao and Li, Xuelong (2017) "Exploring models and data for remote sensing image caption generation", *IEEE Transactions on Geoscience and Remote Sensing*, **5** (4): 2183–2195.
[17] Vaswani, Ashish and Shazeer, Noam and Parmar, Niki and Uszkoreit, Jakob and Jones, Llion and Gomez, Aidan N and Kaiser, Lukasz and Polosukhin, Illia (2017) "Attention is all you need", *Advances in neural information processing systems*, 5998–6008.
[18] Wu, Zifeng and Shen, Chunhua and Van Den Hengel, Anton (2019) "Wider or deeper: Revisiting the resnet model for visual recognition", *Pattern Recognition, Elsevier*, 119–133.
[19] Krizhevsky, Alex and Sutskever, Ilya and Hinton, Geoffrey E (2012) "Imagenet classification with deep convolutional neural networkss", *Advances in neural information processing systems*, 1097–1105.
[20] Simonyan, Karen and Zisserman, Andrew (2014) "Very deep convolutional networks for large-scale image recognition", *arXiv preprint arXiv:1409.155*.