Received: June 29, 2023. Revised: August 6, 2023.

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

Deep Attention Based Dense Net with Visual Switch Added BiLSTM for Caption Generation from Remote Sensing Images

Namdeo Baban Badhe[1]* **Vinayak Ashok Bharadi[1]** **Nupur Giri[2]** **Sujata Alegavi[3]**

[1]*Department of Information Technology, Finolex Academy of Management and Technology,*
P-60, P-60/1, Midc, Mirjole Block, Ratnagiri, Maharashtra-415639, India
[2]*Department of Computer Engineering, Vivekanand Education Society's Institute of Technology,*
Hashu Adwani Memorial Complex, Collector's Colony, Chembur, Mumbai, Maharashtra-400074, India
[3]*Internet of Things Department, Thakur College Engineering and Technology,*
Kandivali - (East), Mumbai, Maharashtra–400101, India
*Corresponding author's Email: namdeobabanbadhe@gmail.com

Abstract: Remote sensing image captioning is the challenging task due to low global information, single feature extraction and lack of detailed image captions. To address these issues, this research proposed a deep attention based DenseNet with visual switch added bidirectional long short-term memory (DADN-BiLSTM) for captioning. In this research, initially the images and captions are collected from captioning dataset to smooth away small structures. After that, a double attention mechanism is applied to DenseNet for capturing weak features and to improve the problem corresponds between image feature and captioning information. At the same time, a clustering-based segmentation is more useful and easier to segment the image as smaller parts to make the access easily. Moreover, a decoder is used to improve the use of captioning context information. Then the proposed system is implemented in PYTHON and the performance is evaluated against existing methods in terms of some relevant evaluation metrics such as, recall-oriented understudy for gisting evaluation, accuracy and bilingual evaluation understudy. Finally, the experimental results achieve higher scores in all evaluation indicators such as 0.8925 BLEU1, 0.8514 BLEU2, 0.8252 BLEU3, 0.8312 BLEU4 and 0.8611 ROUGE score on UCM captions, 0.8532 BLEU1, 0.7912 BLEU2, 0.8351 BLEU3, 0.7215 BLEU4 and 0.8139 ROUGE score on Sydney captions and 0.8125 BLEU1, 0.7501 BLEU2, 0.6812 BLEU3, 0.7254 BLEU4 and 0.8245 ROUGE score on RSICD captions.

Keywords: Remote sensing image, DenseNet, Deep learning, Double attention mechanism, Context information.

1. Introduction

The goal of remote sensing image (RSI) captioning is to automatically create brief words from a high-resolution RSI. [1]. Traditional remote sensing jobs frequently focus on low-level semantic data via image synthesis. [2]. A word level label is given to an RSI by image classification to exchange the low-level semantic information. [3]. An RSI captioning challenge has attracted a lot of attention [4]. The captioning work must be used for a variety of beneficial potential applications, including image retrieval [5-6]. More semantic details about an RSI may be available through the automatic caption generation. [7].

In this research work deep learning (DL) based RSI captioning is proposed to improve the captioning accuracy even for the small object information with the help of classifier. The classification performance is improved by enhancing the image quality and features from the image. Feature extraction and image quality are improved by separate approaches to extract depth and dissimilar features.

The following contributions are focused in this research:

Table 1. Comparison of existing RSI captioning models

Authors	Methods	Pros	Cons
Gajbhiye G. O. et al. [8]	SCAMET and CNN	RSI objects are learned by spatial attention and reducing the entire trained parameter	The small objects are not accurately captured so it suffered from the error rate problems with low computational efficacy and low accuracy. The performance enhances as low values by 3.38%, 13.86%, 16.31% for Sydney-captions, UCM-captions and RSICD respectively
Zhuang S. et al. [9]	Feed Forward Neural Network (FFNN) and SCST based optimization algorithm	Accuracy is improved by introducing grid features.	This technique did not take multi features so it takes high processing time for captioning the image and the BLEU score also reduced Only 10% performance gains in terms of BLEU1 and 0.38 s time taken for testing the image
Hoxha G. et al. [10]	SVM and CNN	Overfitting problems are reduced and computation power is inexpensive	Low global information from feature extraction lowers the captioning scores such as BLEU and ROUGE. So, it takes 35.82 minutes for training and produce 92.48% accuracy.
Wang Q. et al. [11]	GLCM	Word discrimination is increased by considering local features	Failed to align the high-level visual information which affects the detailed description of objects. The average score of BLEU1–BLEU4 has improved from 59.82 to 61.76 with a gain of 1.95
Li Y. et al. [12]	RNN	Provides better context vector to increase the representation of current word states.	Lacked to understand the complex information because it possesses only single features while training the image with captions. It attains 0.8, 0.85 and 0.7729 BLEU 1 score on Sydney-captions, UCM-captions and RSICD respectively
Lu H. et al. [13]	Fuzzy attention based DenseNet-BiLSTM	Improve the captioning by extracting certain features	Complex scenes like small objects are lacked for description which lower the BLUE and ROUGE scores as 0.785 and 0.712

- To expand the variability and accuracy of description production, a vision-language pre-training architecture of DADN- BiLSTM based RSI captioning is proposed.
- The quality of image is affected by various kind of noise; thus, it is needs to smooth away small structure in the image using improved Gaussian rolling guidance filter (IGRGF).
- Double attention-based DenseNet (A^2 DenseNet) is considered with feature extraction approach to connect dissimilar depth features to the attention structure and effectively learning the alignment between image details and words.
- The number of feature maps are reduced to get context information from segmented images.

This research work is organized as follows: Section 2 provides the overview of several existing RSI captioning techniques with RSICD dataset. The proposed methodology is briefly explained in section 3 with the steps of preprocessing, feature extraction (Encoder), segmentation and captioning (Decoder). Section 4 gives the detailed explanation of the experimental result. Finally, the whole research work is concluded with future work in Section 5.

2. Literature survey

Various research works have previously existed in the literature which are based on the RSI captioning. Some of them are reviewed as follows, Gajbhiye, G.O. et al. [8] have presented RSI capturing approach using spatial-channel attention based MEmory-guided transformer (SCAMET) with convolutional neural network (CNN). Another one RSI capturing approach have presented by Zhuang, S. et al. [9] with a combined grid features and transformer-based captioning. In order to improve the accuracy of RSI captioning, self-critical sequence training (SCST) based optimization algorithm was suggested.

Hoxha G. et al. [10] have presented support vector machines (SVM) based decoder for RSI captioning. Again, an RSI captioning technique have suggested by Wang, Q., et.al [11] with a global local captioning model (GLCM) which brings the complete visual significance. Li Y. et al. [12] have presented a recurrent neural network (RNN) with attention-based captioning. Lu H. et al. [13] have presented a DenseNet integrated BiLSTM to improve the extraction of features. The existing models are listed in Table 1.

2.1 Problems of existing approaches

Numerous existing techniques for remote sensing captioning follow the fundamentals of natural image captioning (NIC) and have made some progress when combined with the unique qualities of these images [8-13].

3. Proposed DADN- BiLSTM based RSI captioning

DADN- BiLSTM based RSI captioning is proposed which is shown in Fig. 1.

3.1 Data preprocessing

Initially the input RSIs are taken from remote sensing image captioning datasets (RSICD): RSICD captions, Sydney captions and University of California at Merced (UCM) captions. In data pre-processing, image and text processing are performed to get the data for modelling. Image pre-processing is done with image loading and rearranging into a form of image which is the output of proposed encoder. In text processing lower case conversion for uniformity of text. It delimits the numeric or punctuation mark to remove the unnecessary size of the vocabulary. In IGRGF, two types of filters are integrated to achieve high performance such as rolling guidance filter (RGF) and Gaussian filter by separating the small structure from the large-scale edges through differential operations. The low-frequency information of images is recollected by Gaussian filter where the edge information and tiny structured information are filter out. Moreover, the large-scale edge information is recollected and small structures are eliminated using RGF in the image.

IGRGF is an efficient filtering method for edge-preserving filtering method. The accuracy of image boundary in large-scale is ensured by smoothing small and complex area in the input image which reduces the noise. Initially, microstructures are filtering out by Gaussian filter and restore the edges. The mathematical expression of this filter is shown as Eq. (1) and weight normalization coefficient is calculated in Eq. (2).

$$G(i) = \frac{1}{K_i}\sum_{j\in N_i} exp\left(-\frac{\|i-j\|^2}{2\sigma_S^2}\right) I(j) \quad (1)$$

$$K_i = \sum_{j\in N_i} exp\left(-\frac{\|i-j\|^2}{2\sigma_S^2}\right) \quad (2)$$

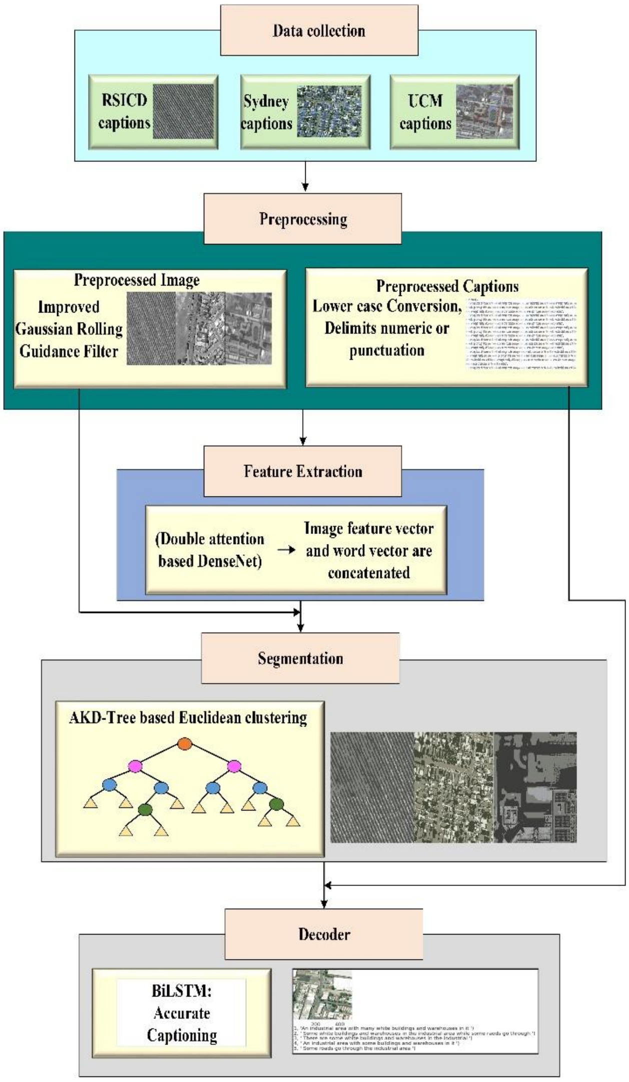

Data collection
RSICD captions
Sydney captions
UCM captions
Preprocessing
Preprocessed Image
Improved Gaussian Rolling Guidance Filter
Preprocessed Captions
Lower case Conversion, Delimits numeric or punctuation
Feature Extraction
(Double attention based DenseNet)
Image feature vector and word vector are concatenated
Segmentation
AKD-Tree based Euclidean clustering
Decoder
BiLSTM: Accurate Captioning

Figure. 1 Work flow of proposed DADN- BiLSTM based RSI captioning

The edge recovery is an iterative process which is mathematically expressed in Eq. (3).

$$J^{t+1}(i) = \frac{1}{k_i}\sum_{j\in N_i} exp\left(-\frac{\|i-j\|^2}{2\sigma_S^2} - \frac{\|J^t(i)-J^t(i)\|^2}{2\sigma_r^2}\right) I(j) \quad (3)$$

This image is suitable for further processing with high outcome thus it given to extract relevant features from large-scale images.

3.2 Feature extraction using A^2 DenseNet

Usually, the image features are extracted and converts as feature map for further processing. Double attention to resolve the problem of variable scale in the RSI. Thus, the convolutional layer is pretrained with RSICD dataset and A^2 DenseNet is used as encoder to extract image features and multi shape information. The N × N pixels of preprocessed image is first extracted and the neighborhood size is determined after that these pixels are given as input to A^2 DenseNet using separate attention mechanism to convolution block. The third layer is used as feature vector for decoder.

Text transformation includes tokenization followed by vectorization of the image descriptions. Initially text feature extraction followed by vectorization of the image description. Then each word in text descriptions is replaced by a numerical value to get word to index and vice versa. The equal sized text vector is obtained by looking the maximum length of the text in which the short text is padded with zero to get equal length to extract context information and cross-channel information. The network weight and similar features at a same location is extracted using two branches. The channel and spatial attentions are employed to achieve deep knowledge of multi-attentive visual, multi-object, multi-scale and multi-shape features.

This network includes convolution block, asymmetric convolution block (ACB), dense block, and transition layer. The number of dense blocks relates to changed dense connection mechanism of DenseNet. The output of the L layer is mathematically expressed as Eq. (4).

$$X_L = H_L([X_0, X_1, .., X_{L-1}]) \qquad (4)$$

The effective feature transfer is achieved by a dense connection mechanism and the gradient disappearance is improved. Overfitting problem is regularized by limiting the parameters of computation. The transition layer that consists of 1 × 1 convolution and average pooling operation are connected with each dense block in DenseNet. Then the image dimension, network complexity and parameters are reduced by the output feature map of number of channels. The block diagram of feature extractor is illustrated in Fig. 2.

The effective feature extraction is done by convolution layer in ACB. The two attention mechanisms are introduced in dense block. The spatial location of particular features is focused by spatial attention mechanism. First, the input feature map performed the pooling operations. Finally, Eq. (5) illustrates the spatial mechanism with convolution operation and activation function. It enhances the spatial information of the model.

$$M(F) = \sigma\left(f\left(AvgPool(F), MaxPool(F)\right)\right) \qquad (5)$$

Channel attention is introduced to take channel information into account by utilizing the relevant information from RSI using feature maps. Every semantic attribute and scale variation learning are employed by the channel attention which is expressed as Eq. (6),

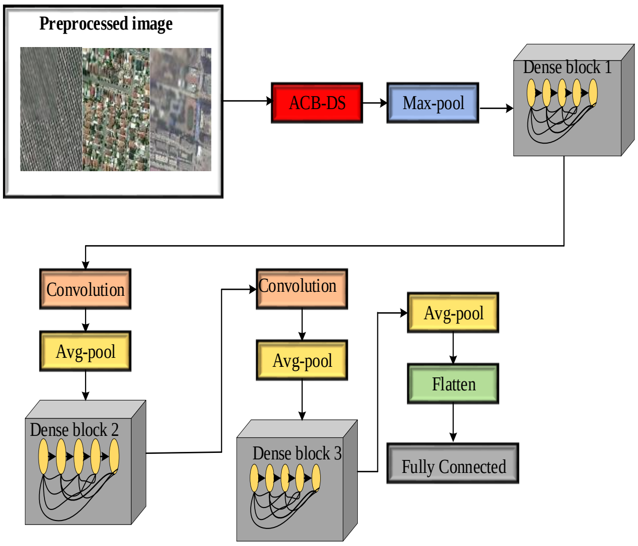


Figure. 2 Structure of A^2 DenseNet

$$C_{avg} = \frac{1}{W\times H}\sum_{i=1}^{W}\sum_{j=1}^{H} R_s(i,j,k) \qquad (6)$$

These values are given to decoder to identify the objects in the image.

The final feature map is used for next captioning stage. Thus, the scale of RSI is varied and cannot deal with the large-scale variation. To fix this problem multi-scale feature fusion (MFF) mechanism is obtained. After getting spatial and channel attention, a denoised feature map of $F'$is created to generate RSI. Thus $F'$is integrated with global vector by global average pooling. The kth channel element is calculated as Eq. (7).

$$M(k) = \frac{1}{H\times W}\sum_{i=0}^{H}\sum_{j=0}^{W} F'(i,j)\; i \in H, j \in W, k \in C_{avg} \qquad (7)$$

Then these features are fused to get MFF as Eqs. (8) and (9).

$$M_{cat} = concat(M_1, M_2, M_3) \qquad (8)$$

$$M_f = f_{cat}(M_{cat}) \qquad (9)$$

The image feature vector and word vector are concatenated by fully connected layer and given to decoder which learns the occurrence current word based on the last predicted and next word in the sequence

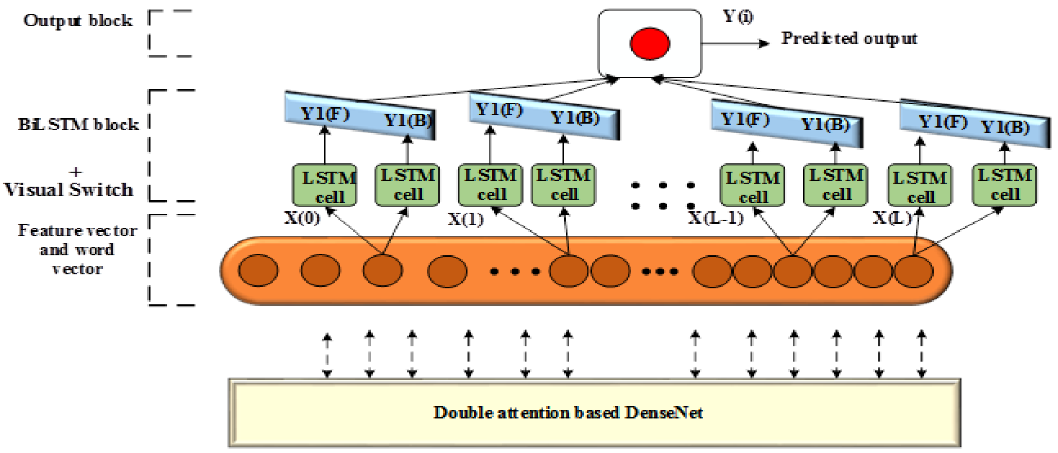


Figure. 3 The architecture of visual switch added decoder

3.3 Segmentation with AKD-Tree based Euclidean clustering

Adaptive K-dimensional tree (AKD-Tree) based Euclidean clustering is introduced in binary search tree-based KD-Tree for the high dimensional searching strategy which prioritize the neighboring point for fast search and set a number of nodes from a group of points by the Euclidean distance. Eq. (10) calculates the Euclidean distance (d_E) in a three-dimensional space.

$$d_E = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2 + (z_1 - z_2)^2} \quad (10)$$

Adaptive clustering distributes the nodes equally at all clusters and segments the image by the information from feature extraction. The adaptive clustering process (A) is expressed in Eq. (11).

$$A = \frac{s}{\omega} \int_{i=1}^{e} \frac{d_E}{f.a} da \quad (11)$$

Then the image features and segmented outcome are giving to decoder for better caption.

3.4 Accurate captioning using BiLSTM

RSI captioning is the recent research area with encoder and decoder architecture with A^2 DenseNet and BiLSTM.

BiLSTM: The decoder model attains the feature vector of image along with the word vector of the original text. The back-propagation path in LSTM can gather the information of previous data. However, BiLSTM can allow to create the forward LSTM layer for backward and forward propagation. Thus, the hidden layers in the BiLSTM can obtain past and future contexts. It is more effective in high frequency data forecasting while compared with traditional LSTM and the computational cost of both layers is same.

Additionally, a visual switch is integrated with the decoder to guide the proposed model to predict whether the captioning words are based on features of image. The visual switch is created at the memory cell of the BiLSTM which stores the current and previous visual information. The value visual switch is computed as Eqs. (12) and (13).

$$a_t = \theta(W_x x_t + W_h h_{t-1}) \quad (12)$$

$$s_t = g_t . tanh\left(\overset{\wedge}{h_t}\right) \quad (13)$$

The forward and reverse BiLSTM network calculation is given in Eq. (14).

$$y_i = s_t\left(w_{01} * h_f + w_{02} * h_b\right) \quad (14)$$

The learned feature vector representation is handled by flatten layer build after these dense block and average pooling. The flattening layer adds compatibility with the next decoder module. The long-term dependency is recorded by the BiLSTM module.

Finally, the overfitting and overwhelming problems are fixed by multiple trials run with 128 neurons. Moreover, the DADN- BiLSTM based RSI captioning method is contrasted with traditional approaches to evaluated its performance in terms of some evaluation parameters. The steps involved in the whole process is listed in Table 2 and the architecture is given in Fig. 3.

4. Experimental analysis and discussion

To make a complete description of RSI, this section discusses various datasets-based different scores to compute the similarity and differences between proposed and traditional methods. This section also takes a quantitative analysis in terms of some evaluation metrics with implemented in PYTHON. The performance metrics such as recall-oriented understudy for gisting evaluation (ROUGE), accuracy and bilingual evaluation understudy (BLEU) are used to compare the proposed modal with the existing methods

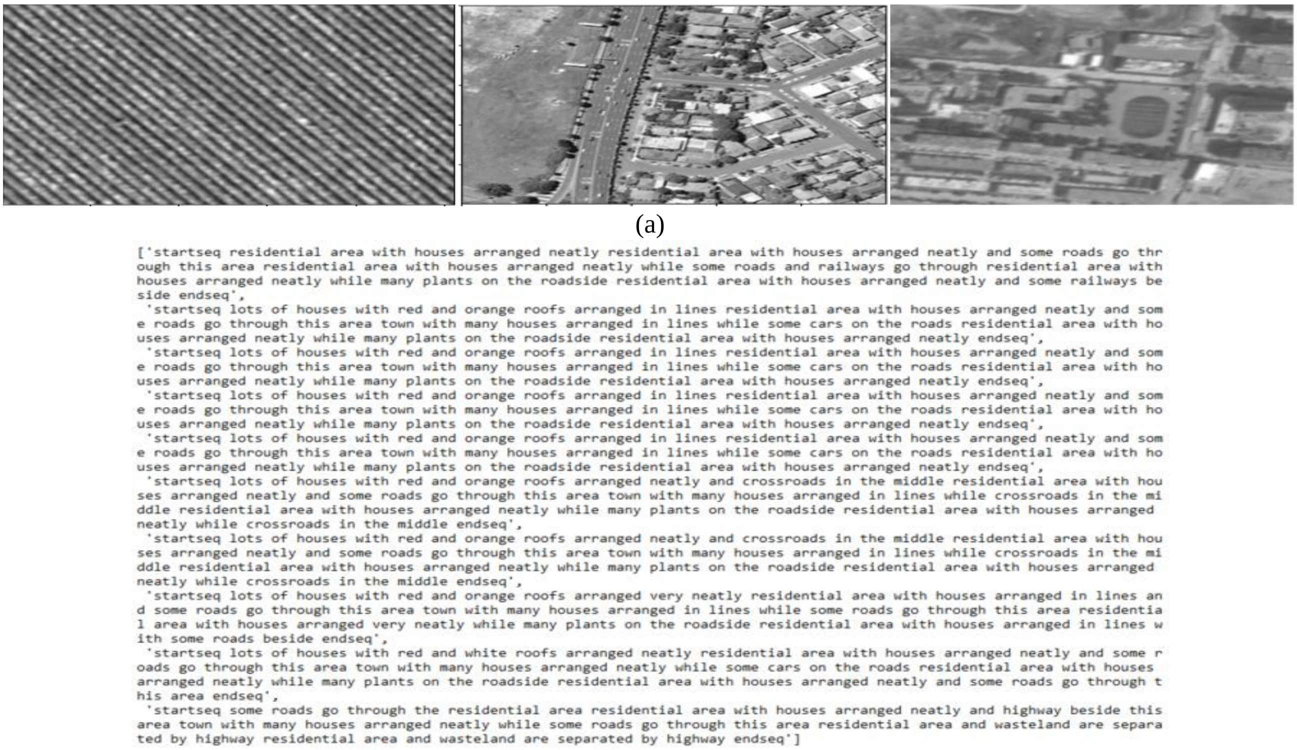
4.1 Dataset description

The selected RSICD captions [19], Sydney captions [19] and UCM captions [19] datasets are used for training (80%), validation (10%) and testing (10%) the RSI captioning process. Their output samples are red-green-blue (RGB) images containing automatic captions. Figure 4 illustrates the input images from datasets.

4.2 Output of preprocessing and segmentation

Fig. 5 shows the outcome of preprocessed image. It separates all small structures of large-scale edges from which the low frequency images are recollected by Gaussian filter.The output image of segmentation is shown in Fig. 6.

Figure. 4 Dataset images



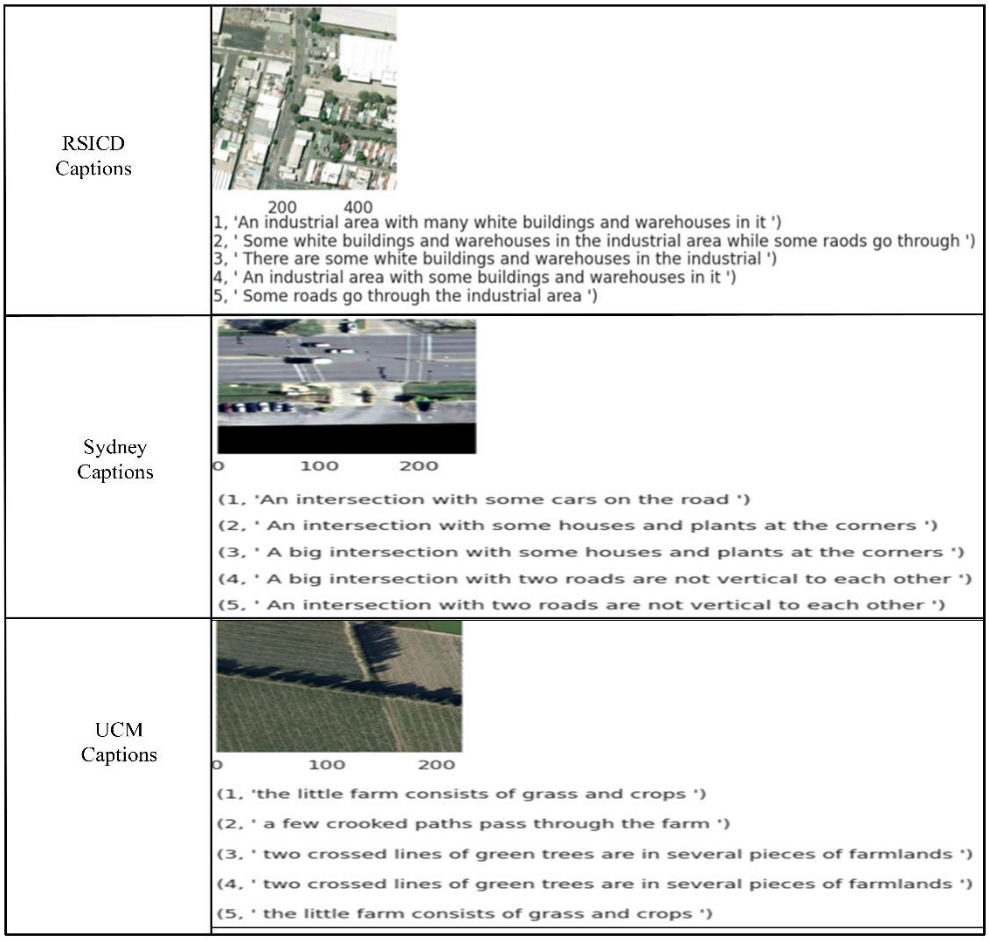
(a)

```
['startseq residential area with houses arranged neatly residential area with houses arranged neatly and some roads go thr
ough this area residential area with houses arranged neatly while some roads and railways go through residential area with
houses arranged neatly while many plants on the roadside residential area with houses arranged neatly and some railways be
side endseq',
 'startseq lots of houses with red and orange roofs arranged in lines residential area with houses arranged neatly and som
e roads go through this area town with many houses arranged in lines while some cars on the roads residential area with ho
uses arranged neatly while many plants on the roadside residential area with houses arranged neatly endseq',
 'startseq lots of houses with red and orange roofs arranged in lines residential area with houses arranged neatly and som
e roads go through this area town with many houses arranged in lines while some cars on the roads residential area with ho
uses arranged neatly while many plants on the roadside residential area with houses arranged neatly endseq',
 'startseq lots of houses with red and orange roofs arranged in lines residential area with houses arranged neatly and som
e roads go through this area town with many houses arranged in lines while some cars on the roads residential area with ho
uses arranged neatly while many plants on the roadside residential area with houses arranged neatly endseq',
 'startseq lots of houses with red and orange roofs arranged in lines residential area with houses arranged neatly and som
e roads go through this area town with many houses arranged in lines while some cars on the roads residential area with ho
uses arranged neatly while many plants on the roadside residential area with houses arranged neatly endseq',
 'startseq lots of houses with red and orange roofs arranged neatly and crossroads in the middle residential area with hou
ses arranged neatly and some roads go through this area town with many houses arranged in lines while crossroads in the mi
ddle residential area with houses arranged neatly while many plants on the roadside residential area with houses arranged
neatly while crossroads in the middle endseq',
 'startseq lots of houses with red and orange roofs arranged neatly and crossroads in the middle residential area with hou
ses arranged neatly and some roads go through this area town with many houses arranged in lines while crossroads in the mi
ddle residential area with houses arranged neatly while many plants on the roadside residential area with houses arranged
neatly while crossroads in the middle endseq',
 'startseq lots of houses with red and orange roofs arranged very neatly residential area with houses arranged in lines an
d some roads go through this area town with many houses arranged in lines while some roads go through this area residentia
l area with houses arranged very neatly while many plants on the roadside residential area with houses arranged in lines w
ith some roads beside endseq',
 'startseq lots of houses with red and white roofs arranged neatly residential area with houses arranged neatly and some r
oads go through this area town with many houses arranged neatly while some cars on the roads residential area with houses
arranged neatly while many plants on the roadside residential area with houses arranged neatly and some roads go through t
his area endseq',
 'startseq some roads go through the residential area residential area with houses arranged neatly and highway beside this
area town with many houses arranged neatly while some roads go through this area residential area and wasteland are separa
ted by highway residential area and wasteland are separated by highway endseq']
```

(b)

Figure. 5: (a) Preprocessed image and (b) preprocessed caption

Figure. 6 Segmented output



RSICD Captions

1, 'An industrial area with many white buildings and warehouses in it ')
2, ' Some white buildings and warehouses in the industrial area while some raods go through ')
3, ' There are some white buildings and warehouses in the industrial ')
4, ' An industrial area with some buildings and warehouses in it ')
5, ' Some roads go through the industrial area ')

Sydney Captions

(1, 'An intersection with some cars on the road ')
(2, ' An intersection with some houses and plants at the corners ')
(3, ' A big intersection with some houses and plants at the corners ')
(4, ' A big intersection with two roads are not vertical to each other ')
(5, ' An intersection with two roads are not vertical to each other ')

UCM Captions

(1, 'the little farm consists of grass and crops ')
(2, ' a few crooked paths pass through the farm ')
(3, ' two crossed lines of green trees are in several pieces of farmlands ')
(4, ' two crossed lines of green trees are in several pieces of farmlands ')
(5, ' the little farm consists of grass and crops ')

Figure. 7 RSI captioning of dataset images

The ground and individual structured segmented objects in the image are mentioned with different colors and numbers through duplicate Euclidean distance. Fig. 7 shows the prediction outcome of generating captions by BiLSTM. The encoder extracts the features and the text quality depends on segmentation output and relevant features. Thus, the different scale features are extracted for the full utilization of image features which is used to enhance the mapping of feature to learn the complete contextual information with word and feature vectors.

4.3 Comparison of multimodal method on different captions

This section evaluates RNN and LSTM methods with the proposed feature extractor [11]. Table 2 illustrates the multimodal comparison with the proposed feature extractor where "- "indicates that the corresponding score is not available in that research article. In existing feature extraction methods, spatial details might be lost or downplayed, leading to a reduction in the quality and accuracy of the generated captions also some methods may struggle to effectively represent and incorporate temporal information, resulting in captions that fail to reflect the dynamic nature of the scenes and high-dimensional feature space, which can increase computational costs and pose challenges for subsequent processing and modeling. Understanding the interactions between different objects and structures in the scene is crucial for generating meaningful and accurate captions.

All the listed values show that the score is high on Sydney, UCM and RSICD captions which concludes that the proposed method attains high performance while extracting the image and text features for captioning. The proposed A^2 DenseNet accurately extract the global and local features along with multiscale features. This gives 0.8125, 0.8532 and 0.8925 BLEU scores on RSICD dataset.

4.4 Parameter comparison with existing methods

The proposed approach is compared with other existing methods in RSI captioning on three captions such as Sydney, UCM and RSICD dataset. In Table 3 "-"indicates that the corresponding score is not available in that research article.

Table 2. Multimodal comparison on UCM, Sydney and RSICD captions

		BLEU 1	BLEU 2	BLEU 3	BLEU 4	ROUGE
RNN	Scale Invariant Feature Transform (SIFT)	0.573	1.443	0.379	0.338	0.533
	Bag of Words (BOW)	0.410	0.224	0.145	0.109	0.343
	Fisher Vector (FV)	0.390	0.460	0.396	0.354	0.545
	Vector of Locally Aggregated Descriptors (VLAD)	0.631	0.519	0.460	0.420	0.58
LSTM	SIFT	0.551	0.416	0.348	0.304	0.523
	BOW	0.391	0.187	0.108	0.070	0.331
	FV	0.589	0.466	0.407	0.368	0.55
	VLAD	0.701	0.608	0349	0.503	0.651
DADN- BiLSTM (UCM captions)		0.8925	0.8514	0.8252	0.8312	0.8611
RNN	SIFT	0.588	0.481	0.426	0.389	0.539
	BOW	0.531	0.407	0.331	0.278	0.492
	FV	0.605	0.491	0.42	0.378	0.554
	VLAD	0.565	0.451	0.380	0.327	0.527
LSTM	SIFT	0.579	0.477	0.418	0.37	0.536
	BOW	0.531	0.407	0.331	0.278	0.492
	FV	0.633	0.533	0.473	0.430	0.579
	VLAD	0.588	0.481	0.426	0.389	0.539
DADN- BiLSTM (Sydney captions)		0.8532	0.7912	0.8351	0.7215	0.8139
RNN	SIFT	0.476	0.282	0.196	0.145	0.399
	BOW	0.440	0238	0.151	0.104	0.360
	FV	0.485	0.303	0.218	0.167	0.417
	VLAD	0.493	0.309	0.220	0.167	0.424
LSTM	SIFT	0.485	0.303	0.218	0.167	0.417
	BOW	0.481	0.290	0.204	0.153	0.399
	FV	0.434	0.245	0.163	0.117	0.381
	VLAD	0.500	0.319	0.23	0.177	0.433
DADN- BiLSTM (RSICD captions)		0.8125	0.7501	0.6812	0.7254	0.8245

Table 3. Evaluation score of proposed method with existing methods

Model	**BLEU1**	**BLEU2**	**BLEU3**	**BLEU4**	**ROUGE**
Sydney captions					
SCAMET [8]	0.8072	0.7136	0.6431	0.5846	0.7320
SCST [9]	0.810	0.732	0.650	0.571	0.749
SVM-CNN [10]	0.7547	0.6711	0.5970	0.5308	0.6746
GLCM [11]	0.8041	0.7305	0.6745	0.6259	0.6965
Attention based RNN [12]	0.8000	0.7217	0.6531	0.5909	0.7218
GoogleNet [11]	0.7909	0.7221	0.6546	0.5983	0.6950
DADN- BiLSTM	0.8532	0.7912	0.8351	0.7215	0.8139
UCM captions					
SCAMET [8]	0.8460	0.7772	0.7262	0.6812	0.8166
SCST [9]	0.834	0.772	0.718	0.673	0.770
SVM-CNN [10]	0.7653	0.6947	0.6417	0.5942	0.6877
GLCM [11]	0.8182	0.7540	0.6986	0.6468	0.7524
GoogleNet [11]	0.8077	0.7448	0.6930	0.6446	0.7356
Attention based RNN [12]	0.8518	0.7925	0.7432	0.6976	0.8072
DADN- BiLSTM	0.8925	0.8514	0.8252	0.8312	0.8611
RSICD captions					
SCAMET [8]	0.7693	0.6309	0.5352	0.4611	0.6979
SCST [9]	0.774	0.668	0.581	0.510	0.678
SVM-CNN [10]	0.5999	0.4347	0.3355	0.2689	0.4557
GLCM [11]	0.7767	0.6492	0.5642	0.4937	0.6779
GoogleNet [11]	0.7588	0.6323	0.5375	0.4640	0.6732
Attention based RNN [12]	0.7729	0.6651	0.5782	0.5062	0.6691
DADN- BiLSTM	0.8125	0.7501	0.6812	0.7254	0.8245

The SCAMET [8] system faces challenges in accurately capturing small objects, resulting in higher error rates and reduced efficiency with low accuracy. Due to SCST [9] systems' inability to handle multiple features, the image captioning process is time-consuming, leading to decreased BLEU scores with only 10% performance gain for BLEU1. Testing an image takes 0.38 seconds in SVM-CNN [10]. The lack of global information from feature extraction adversely affects captioning scores. The system still struggles to align high-level visual information, impacting detailed object descriptions in GLCM [11]. The Attention based RNN [12] has limitations that includes a lack of understanding complex information since it relies on single features during image training.

Applying BiLSTM with visual switch gives better accuracy and provides coherent sentences. This is because the visual switch again predict that the generated sentences are depend on image features. The MFF is applied in proposed system which also extracts the multiscale features which will increase the performance in terms of 0.8125 BLEU1, 0.7501 BLEU2, 0.6812 BLEU3, 0.7254 BLEU4 and 0.8245 ROUGE scores on RSICD captions, 0.8925 BLEU1, 0.8514 BLEU2, 0.8252 BLEU3, 0.8312 BLEU4 and 0.8611 ROUGE scores on UCM captions and 0.8532 BLEU1, 0.7912 BLEU2, 0.8351 BLEU3, 0.7215 BLEU4 and 0.8139 ROUGE on Sydney captions. Thus the proposed method attains high degree of accuracy in all scores for captioning. The preprocessed output gives the quality image and corresponding captions for image text.

5. Conclusion

In this research work, DADN-BiLSTM based method was successfully implemented with the full use of global and multiscale features for accurate captioning. The structured images and captions were collected with improved filtering approach to segment the image by adaptively distributes the nodes equally at all clusters. Moreover, visual switch added BiLSTM improved the captioning to predict whether the captioning words are based on features of image. Finally, the proposed approach was implemented in PYTHON and evaluated against various existing approaches to find the performance and computational efficiency of the proposed system. Thus, the proposed model achieves higher performance of 0.8125 BLEU1, 0.8245 ROUGE scores on RSICD captions, 0.8252 BLEU3, 0.8545 ROUGE scores on UCM captions and 0.8532 BLEU1, 0.8139 ROUGE scores on Sydney captions.

In future, RSI captioning will motivate the generation of new sentences and high-quality text annotations. Also, it will concentrate the captioning technique with some other evaluation metrices. Next, it will optimize the segmentation and large visual relation to improve the captions for the complex images. Moreover, the accurate feature extraction strategy will be motivated in further researches.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Conflict of interest

The authors declare that they have no potential conflict of interest.

Authors' contribution

Namdeo Baban Badhe proposed the idea. Vinayak Ashok Bharadi took the lead in writing the manuscript. Nupur Giri and Sujata Alegavi supervised the findings and all authors discussed the results and contributed to the final manuscript.

Notations

Notations	Descriptions
$I(j)$	Input image
$G(i)$	Pre-processed output image
$M(F)$	Spatial attention mechanism
C_{avg}	Channel wise spatial feature
M_{cat}	Concatenation of large-scale features
$F_1, F_2, F_3.$	Large scale features
M_1, M_2, M_3	Global features
d_E	Euclidean distance
A	Adaptive clustering
ω	Node permission value in the adaptive clustering
a_t	Previous information
θ	Current information
W_x, W_h	Weight values of the memory cell
$\overset{\wedge}{h_t}$	Memory cell
s_t	Visual switch
σ_s	Spatial weight
σ_s	Representation of standard deviation of gaussian filter
$N(i)$	Neighborhood centered on pixels
i, j	Pixels
K_i	Weight normalization coefficient
σ_r	Spatial weight
$J^{t+1}(i)$	Output result of the t^{th} iteration
$H_L(.)$	Nonlinear transformation function
X_L	Output of the L layer

X_0, X_1, X_{L-1}	Output of each layer in the dense block
$AvgPool(F), MaxPool$	Average and maximum feature map of pooling layer
$W \times H$	Feature map of input
R_s	Visual feature map
$H \times W$	Spatial location
s	Feature vector in the segmentation
e, i	Upper and lower boundary values of feature extraction
f	Qualitative indicator for clustering
a	Average value of extracted data
W_x and W_h	Weight values of the memory cell
y_i	Final output of LSTM
h_f, h_b	Output of forward and reverse network
w_{01},w_{02}	Weight values

References

[1] R. Zhao, Z. Shi, and Z. Zou, “High-Resolution Remote Sensing Image Captioning Based on Structured Attention”, *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 60, pp. 1–14, 2022.

[2] N. Murali and A. P. Shanthi, *Remote Sensing Image Captioning via Multilevel Attention-Based Visual Question Answering, In: S. Roy, D. Sinwar, T. Perumal, A. Slowik, J.M.R.S. Tavares, (eds.) Innovations in Computational Intelligence and Computer Vision*, Advances in Intelligent Systems and Computing, Vol. 1424. Springer, Singapore, 2021.

[3] U. Zia, M. Mohsin Riaz, and A. Ghafoor, “Transforming Remote Sensing Images to Textual Descriptions”, *International Journal of Applied Earth Observation and Geoinformation*, Vol. 108, p. 102741, 2022.

[4] G. Hoxha, S. Chouaf, F. Melgani, and Y. Smara, “Change captioning: A New Paradigm For Multitemporal Remote Sensing Image Analysis”, *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 60, pp. 1–14, 2022.

[5] R. Ramos and B. Martins, “Using Neural Encoder-Decoder Models with Continuous Outputs for Remote Sensing Image Captioning”, *IEEE Access*, Vol. 10, pp. 24852–24863, 2022.

[6] H. Zhou, X. Du, L. Xia, and S. Li, “Self-learning for Few-Shot Remote Sensing Image Captioning”, *Remote Sensing*, Vol. 14, No. 18, p. 4606, 2022.

[7] G. Wang, B. Li, T. Zhang, and S. Zhang, “A Network Combining a Transformer and a Convolutional Neural Network for Remote Sensing Image Change Detection”, *Remote Sensing*, Vol. 14, No. 9, p. 2228, 2022.

[8] G. O. Gajbhiye and A. V. Nandedkar, “Generating the Captions for Remote Sensing Images: A Spatial-Channel Attention-Based Memory-Guided Transformer Approach”, *Engineering Applications of Artificial Intelligence*, Vol. 114, p. 105076, 2022.

[9] S. Zhuang, P. Wang, G. Wang, D. Wang, J. Chen, and F. Gao, "Improving Remote Sensing Image Captioning by Combining Grid Features and Transformer", *IEEE Geoscience and Remote Sensing Letters*, Vol. 19, pp. 1-5, 2022, Art no. 6504905.

[10] G. Hoxha and F. Melgani, “A Novel SVM-based Decoder for Remote Sensing Image Captioning”, *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 60, pp. 1–14, 2022.

[11] Q. Wang, W. Huang, X. Zhang, and X. Li, “GLCM: Global–Local Captioning Model For Remote Sensing Image Captioning”, *IEEE Transactions on Cybernetics*, pp. 1–13, 2022.

[12] Y. Li, X. Zhang, J. Gu, C. Li, X. Wang, X. Tang, and L. Jiao, “Recurrent attention and Semantic Gate for Remote Sensing Image Captioning”, *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 60, pp. 1–16, 2022.

[13] H. Lu, R. Yang, Z. Deng, Y. Zhang, G. Gao, and R. Lan, “Chinese Image Captioning via Fuzzy Attention-Based DenseNet-BILSTM”, *ACM Transactions on Multimedia Computing, Communications, and Applications*, Vol. 17, No. 1s, pp. 1–18, 2021.

[14] R. Ramos and B. Martins, “Using neural encoder-decoder models with continuous outputs for remote sensing image captioning”, *IEEE Access*, Vol. 10, pp. 24852–24863, 2022.

[15] R. Ramos and B. Martins, “Using Neural Encoder-Decoder Models with Continuous Outputs for Remote Sensing Image Captioning”, *IEEE Access*, Vol. 10, pp. 24852–24863, 2022.

[16] Q. Wang, W. Huang, X. Zhang, and X. Li, “Word–Sentence Framework for Remote Sensing Image Captioning”, *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 59, No. 12, pp. 10532–10543, 2021.

[17] S. C. Kumar, M. Hemalatha, S. B. Narayan, and P. Nandhini, “Region Driven Remote Sensing Image Captioning”, *Procedia Computer Science*, Vol. 165, pp. 32–40, 2019. doi: 10.1016/j.procs.2020.01.067

[18] S. Zaoad, M. M. R. Mannan, A. B. Mandol, M. Rahman, A. Islam, and M. Rahman, “An Attention-Based Hybrid Deep Learning Approach for Bengali Video Captioning”, *Journal of King Saud University - Computer and*

Information Sciences, Vol. 35, No. 1, pp. 257–269, 2023.

[19] https://drive.google.com/drive/folders/1BV6qJufYllZ7oR0EEv-UMYt7bhYOFFc1