Deformable and Structural Representative Network for Remote Sensing Image Captioning

Jaya Sharma[1], Peketi Divya[2], C. Vishnu[1], C. Linga Reddy[4], B. H Sekhar[3], C. Krishna Mohan[1]
[1]Department of Computer Science and Engineering, Indian Institute of Technology Hyderabad, Hyderabad, India.
[2]Department of Artificial Intelligence, Indian Institute of Technology Hyderabad, Hyderabad, India.
[3]Department of Computer Science, Mangalore University, Karnataka, India.
[4]Department of Information and Communication Technology, UiA, Campus Grimstad, Norway.

{cs18m19p100002@iith.ac.in, ai21resch01001@iith.ac.in, cs16m18p000001@iith.ac.in, linga.cenkeramaddi@uia.no, bhshekar@gmail.com, ckm@cse.iith.ac.in}

Keywords: Deformable network, Contextual network, Structural representative network, Attention mechanism, Multi-level LSTM,.Remote sensing image captioning

Abstract: Remote sensing image captioning has greater significance in image understanding that generates textual description of aerial images automatically. Majority of the existing architectures work within the framework of encoder-decoder structure. However, it is noted that the existing encoder-decoder based methods for remote sensing image captioning avoid fine-grained structural representation of objects and lack deep encoding representation of an image. In this paper, we propose a novel structural representative network for capturing fine-grained structures of remote sensing imagery to produce fine grained captions. Initially, a deformable network has been incorporated on intermediate layers of convolutional neural network to take out spatially invariant features from an image. Secondly, a contextual network is incorporated in the last layers of the proposed network for producing multi-level contextual features. In order to extract dense contextual features, an attention mechanism is accomplished in contextual networks. Thus, the holistic representations of aerial images are obtained through a structural representative network by combining spatial and contextual features. Further, features from the structural representative network are provided to multi-level decoders for generating spatially semantic meaningful captions. The textual descriptions obtained due to our proposed approach is demonstrated on two standard datasets, namely, the Sydney-Captions dataset and the UCM-Captions dataset. The comparative analysis is made with recently proposed approaches to exhibit the performance of the proposed approach and hence argue that the proposed approach is more suitable for remote sensing image captioning tasks.

1 INTRODUCTION

Due to technological advancements in aerial imager understanding systems, it is possible by the satellites an launch vehicles to provide the finer details of the earths surface. However, the amount of remotely sensed imag data generated by satellites and launch vehicles is very large, making it difficult for both researchers and users to access, store, and observe relevant data from a large number of details. Remote sensing image captioning (RSIC) is a probabilistic approach that depends on the attributes and visual features of the image and it helps in providing meaningful descriptions about remotely sensed images. The main goal of RSIC is that the generated captions must focus on the relationship between the scene and the object of an image. It is applied in most of the recent applications, including

[1] https://orcid.org/0000-0000-0000-0000
[2] https://orcid.org/0000-0000-0000-0000

image interpretation and understanding [Wang B, 2020], text search in an image [Gu J, 2018], detail generation [Xu K, 2015], robotic vision [Szegedy C, 2016], content search [Chung YN, 2015] etc. Several RSIC methods [Ramos, 2022 & Wang, 2022] in the literature explore encoder-decoder framework to generate meaningful captions. These frameworks generate sequences of words using recurrent neural networks (RNNs) that depend on the features of the remote sensing images captured by the convolutional neural networks (CNNs). Recently, Gu et al. [Ma C, 2018] proposed an approach in which multiple LSTM decoders are incorporated into a multi level coarse to fine generative network for producing fine grained descriptions. Later, a few research works employed the attention mechanism [Wei H, 2019] in the encoder and decoder of the models to focus on essential parts of an image. However, these frameworks are not efficient in generating meaningful captions due to their failure in capturing the structural representations of remotely sensed images (RSI). The major challenges pertaining to current approaches are (i) inefficient spatial and structural representations of RSI and (ii) the use of a single-level caption decoder, which leads to inefficient generation of fine-grained meaningful captions. Therefore, we propose a structural representative network (SRN) that addresses these limitations.

Our proposed SRN extracts both spatial and semantic information from initial and final layers of CNN through deformable and contextual networks. Deformable networks are incorporated into initial layers of CNN as it possesses capability of handling geometric transformations, thus improving the transformation capability of the model. However, CNNs do not handle these transformations because of fixed size kernels of convolution and max pool operations. Robust way of representing different scaled objects is possible through a deformable network in remote sensing images. An SRN is incorporated on top of the final layers of CNN to capture fine grained structural representations of RSI since CNN fails in preserving multi scale features and capturing boundary details of an object due to repeated stride and pooling operations [Chen LC, 2018]. The proposed SRN captures structural representations of an RSI by applying dilated convolutions parallely at different dilation rates in various fields of views. Finally, these spatial and semantic features are given to multi level decoder for generating meaningful captions of an RSI. The major contributions of this work are as follows:

- We propose a structural representative network (SRN) in order to encode guided contextual information for generating meaningful captions.
- A novel structural representation and deformable networks are incorporated at the encoder of our encoder decoder framework for capturing dense multi scale and spatial invariant features from RSI for meaningful caption generation
- Accuracy of the proposed network is evaluated on two RSIC datasets, namely, Sydney captions and UCM captions. Our model outperforms state of the art methods in generating semantically meaningful captions by learning a rich structural representation of an RSI.

The remainder of the paper is organised as follows. A detailed description of the existing works on RSIC is provided in Section 2. Our proposal method is presented in Section 3. The implementation details, experimental results, comparison with state-of-the-art methods are shown in Section 4. Finally, Section 5 provides the conclusion.

2 Existing Works

This section reviews the state of the art methods and various statistical learning algorithms used in structural investigation and semantic retrieval of RSIC.

2.1 Image Captioning in Remote Sensing Imagery

Currently, deep learning models are extensively explored and applied in image captioning tasks where input can be camera based or remotely sensed images. Most of the image captioning tasks [zhang, 2017] utilise encoder decoder frameworks where CNNs are used at encoder to capture the meaningful features and RNNs at decoder for providing textual description of an RSI. Similarly, we have seen many promising approaches [lu, 2019], [Zhao R, 2021], [Wang Q, 2020], [Ma X, 2020] to generate descriptions for RSI. It is also noted that template-based technique was proposed by Shi et al [Shi Z, 2017] where FCN and CNN are used to capture the image content. Here, the generated words from the decoder are sequenced into complete sentences with stable templates. Further, Lu et al [Lu X, 2017] introduced RSI datasets such as Sydney-Captions and the UCM-Captions datasets for image captioning tasks, and conducted a series of evaluations and observed increase in the performance with these datasets by applying a soft attention mechanism. Later, the attribute-attribute model was proposed by Zhang et. al [Zhang X, 2019] to generate a huge amount of salient features by capturing the core attributes to reweight the features of an image. It is proved that the strength of the model is improved with this attribute data for

generating meaningful sentences. Marker-driven attention mechanism was introduced by Zhang et al [Zhang Z, 2019] for generating meaningful captions where the details of the marker are utilised in the attention computation to assign weights in the attention layer. In this approach, label data from RSI is filtered out and more salient features are provided in each decoding step. Another notable work by Sumbul et al [Sumbul G, 2020] considered novel visual alignment loss to capture important features from an RSI for the training process.

2.2 Visual feature encoding methods

Standard pipeline with combination of CNN and RNN was exploited by Ma et al. [Jégou H, 2011] for capturing rich features of an image using deep CNN to improve the performance of visual tracking. Firstly, the model combines all five true captions into single caption and then adaptive weighting strategy applied dynamically considering the predicted words in the standard and summary time steps. Fusion method combines semantic and coarse details to achieve spatial visual features of an image. ExFuse was another fusion method proposed for segmentation tasks where semantic features are combined with low level features of an image and high spatial resolution features are combined with high level features. Later, spatial transformer networks (STN) [Jaderberg M, 2015] was introduced to deal with spatial transformations of an image. This approach can even include spatial manipulations into the training data. Further, dilated convolutions are applied in CNNs to obtain multi-level structural representations of an image. With the exponential increase in size of the receptive fields of dilated convolutions, rich structural representations are captured [Dai J, 2017]. Then, the spatial pyramid pooling component [Chen LC, 2018] was introduced to study dilated convolutions with different dilation rates in multiple fields of views. Deformable networks [Dai J, 2017] was introduced with inspiration from STNs. Incorporating deformable networks improves the performance of the model by generating dense predictions compared to STNs. Recently, several works [Lu Y, 2020], [Jaderberg M, 2015] handled geometric transformations and multi-scale variations by incorporating deformable networks in model architectures.

3 Proposed Methodology

This section consists of encoding and decoding phase details. Convolutional neural network (CNN) is employed in the encoding phase to encode the visual depiction of an RSIs followed by a decoding phase where multi level long short-term memory (LSTM) is utilised to decode the visual attributes and produce a series of words. Subsequently, the structural representative network and deformable networks are designed for RSI captioning. The framework of the proposed methodology is presented in figure 1.

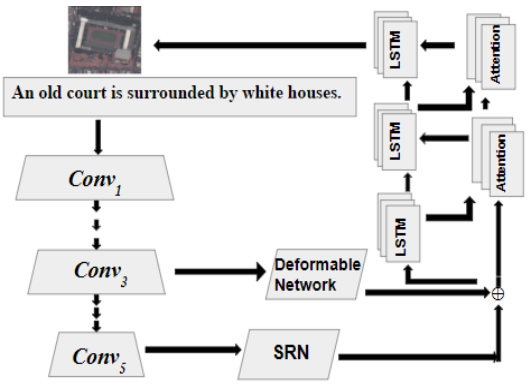


Figure 1. The overview of the proposed method.

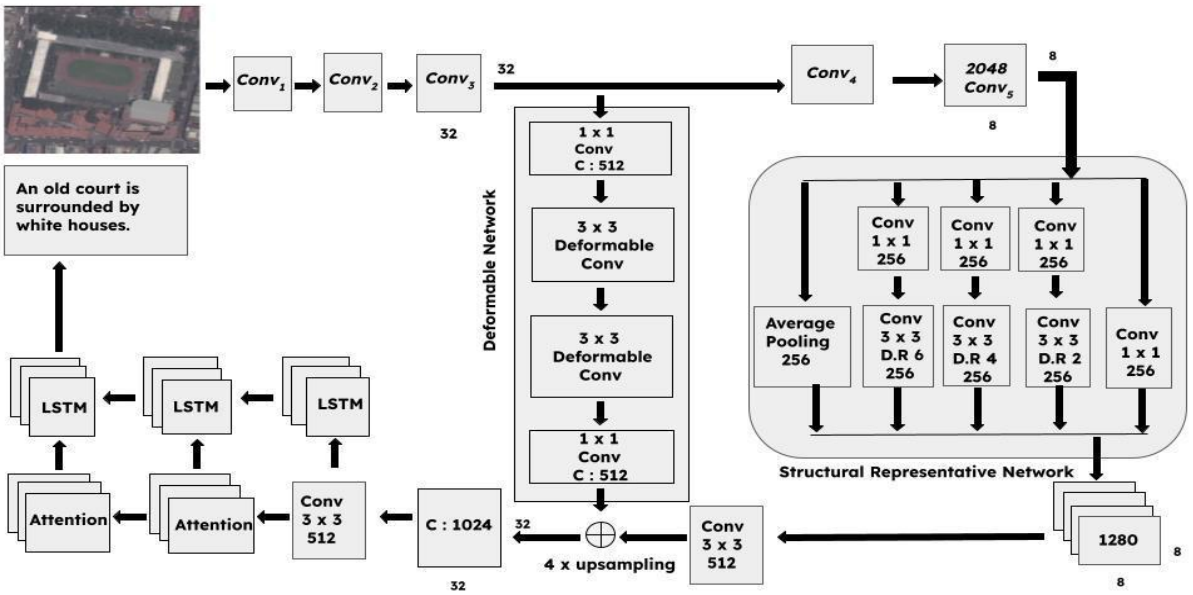
3.1 Visual encoder

Visual encoder network helps in encoding the features where the multiscale structural representations and spatially transformed features of an image are encoded by processing the remote sensing image through distinct network components namely the backbone network, deformable network, multi scale structural representative network, and feature fusion.

3.1.1 Backbone network

A pre-trained ResNet [He K, 2016] is used as a backbone network and hence to retrieve the features for providing visual descriptions of an RSI through the first five layers, namely, Conv1, Conv2, Conv3, Conv4, and Conv5. Each Conv layer contains different bottleneck layers. Generally, initial layers of the CNNs hold small object details in an image but fail in capturing the semantic details whereas final layers can capture semantic details, lacking spatial information of the objects. In our framework (figure 1), fully connected layers of ResNet are removed to utilise spatial and structural representations of conv layers. We incorporate 32×32 with 512 channels spatial resolution at conv3 of the backbone ResNet network after extracting the spatial features from it. Further, semantic details are obtained through conv5 layer with 8x8 resolution and 2048 channels. Deformable network and SRN is incorporated on top of initial and final layers of backbone network to capture multi level structural representations of RSI.

Figure 2 : Framework of proposed methodology



3.1.2 Deformable network

The CNN is a broadly analysed model for image captioning, yet its accomplishment is bounded by the absence of capacity to deal with geometrical changes. Earlier, the CNN networks acquired spatially invariant information by usage of enormous approaches, augmentation methods, and hand crafted methods, for example, max-pooling or scale invariant element changes using SIFT. But, hand crafted features cannot deal with unspecified geometric transformations [Si H, 2019] as it can learn only fixed and known changes. Further, these models and hand-crafted approaches are troublesome, intricate, inaccessible, and require costly preparation for excessively complex changes. Furthermore, the same activation units are produced by fixing the kernel and max pooling receptive field sizes by skipping out the fact that the different locations map with different scaled objects as well as deformations [Si H, 2019]. To overcome these problems, we have introduced deformable convs on Conv3 of backbone network as the initial layers are not genuinely invariant to enormous changes of the input data when compared to final layer attributes (Conv5). The standard convolution produces the output feature map L utilising grid ℊ over the input feature map F as

$$L(A_0) = \sum \wp(A_n) \cdot F(A_0 + A_n + \Delta A_n), \quad A_n \in ℊ \tag{1}$$

Where A signifies random position ($F(A_0 + A_n + \Delta A_n)$ and $\wp$, A_0 signifies the weight values and location on L . A_n defined on sampled grid (ℊ). Thus, the offsets are augmented using deformable convolutions $\Delta A_m \mid m = 1, \ldots, M$, $M = \text{mod } ℊ$. In deformable convolutions, the sampling is done at irregular and offset locations ($A_n + \Delta A_n$). As shown in Eq 2, we apply bilinear interpolation ß on Eq 1 since ΔA_n is fractional and all spatial locations in F are enumerated by S.

$$F(p) = \sum (ß(S,P) \cdot F(S)) \tag{2}$$

In the deformable convolution module, 2D offsets are sequentially added at grid sampling points in standard convolutions. This offset is learned from the previous convolution layer and determines the deformation of the input feature. Simple backward propagation can be used to train the end-to-end deformable convolution module between layers. This work incorporates two deformable convolutions in between two standard bottleneck layers..

3.1.3 Multi-scale structural representation network

Major works in image captioning [Vinyals O, 2015], [Xu K, 2015], [Chen L, 2017], [Wei H, 2019] utilise the semantic details from the final layers of the backbone

network. But, boundary specific details of the objects are reduced due to application of a large number of pooling and convolutions operations. To overcome this problem, we have employed SRN on top of conv5 feature map for capturing structural representations of the RSI by applying multiple dilated convolutions with different scales parallely. The SRN helps in mitigating the number of learnable parameters along with computation time by managing the receptive field size of the input feature map instead of increasing the filter's field of view. Thus, greater structural details are obtained through SRN where it segments the objects at various scales at every layer through parallel dilated convolutions at different dilation rates. More formally, given a 2D signal for every location *L* on *m* which is the output attribute map and ৩ as the kernel matrix with weights, the dilated convolutions are utilised on the input attribute map F as

$$L[q]=\sum F[q+D\cdot \text{ㅗ}]\text{৩}[\text{ㅗ}] \quad (3)$$

Here, for each location *L,* of q as the output feature map and ৩ as weight kernel matrix, D implies the dilation rate at which the sampling of the input feature map is done. D = 1 is the dilated convolution's special case, where it denotes the ideal convolution. The filter's field-of-view changes flexibility with dilation rates.

3.1.4 Feature fusion

The spatially transformed representations of RSI obtained from deformable network at Conv3 of backbone network and structural representations obtained due to feeding Conv5 features of backbone network to SRN networks are fused in the proposed method. Before concatenating, bi-linear interpolation technique is applied to upsample structural features to spatial features. Then, concatenation of these features are done through stacking up features one another and 3x3 convolutions are applied before sending to the decoder.

3.2 Decoding Module

Almost all recent works used a one stage decoder module for generating captions of an RSI. However, these works failed in obtaining meaningful captions as there is minimal transitional supervision. To overcome this challenge, we have utilised multistage caption decoder framework [Gu J, 2018] for a coarse to fine grained caption generation. It also focuses on the problem of vanishing gradients that arose because of the coarse-to-fine multi-stage caption decoding module. As shown in Figure 1, a three-stacked long-short term memory network, where stage one of the LSTM decoder generates coarse to fine grained RSI details, and the succeeding LSTM decoder produces the fine-grained details. At every stage, preceding decoder hidden vectors and attention weights are given to the next LSTM decoder to produce more precise captions.

A stage-wise details of multi-stage decoder are provided as follows: LSTM network ($LSTM_C$) learns encoded details of an image by utilising the visual encoder. At every interval ㅗ, the details of preceding words, the visual depiction of an image, and the LSTM network's previous hidden states are provided to $LSTM_C$ to produce the caption as given by

$$C^0{}_{\text{ㅗ}}, H^0{}_{\text{ㅗ}} = LST\,M_C(H^0{}_{\text{ㅗ}-1}, X^0{}_{\text{ㅗ}}, W_{\text{ㅗ}-1}),$$

$$X^0{}_{\text{ㅗ}} = [f(Z); H^{Nf}{}_{\text{ㅗ}-1}], \quad (4)$$

where the hidden states are $H^0{}_{\text{ㅗ}-1}$ and $H^{Nf}{}_{\text{ㅗ}-1}$, the cell state is $C^0{}_{\text{ㅗ}}$, the preceding word is $W_{\text{ㅗ}-1}$, X denotes (X = 0 for $LSTM_C$ and X & $d_{\text{ㅗ}}$ = 1 for fine decoders ($LSTM_f$)). The total number of fine stages are indicated by N_f, and the mean pool visual encoder features are denoted by f(Z). Further, utilising the attention weights $\alpha^{X-1}{}_{\text{ㅗ}}$, fine stage decoders, visual details, and preceding words are precisely captioned as

$$C^X{}_{\text{ㅗ}}, H^X{}_{\text{ㅗ}} = LST\,M_f(H^X{}_{\text{ㅗ}-1}, X^X{}_{\text{ㅗ}}, W_{\text{ㅗ}-1}),$$

$$X^X{}_{\text{ㅗ}} = [d(Z, \alpha^{X-1}{}_{\text{ㅗ}}, H^{X-1}{}_{\text{ㅗ}}, H^{X-1}{}_{\text{ㅗ}}], \quad (5)$$

Here, the function of spatial attention (d(.)) is to produce attention-guided visual details. On accomplishing attentive attributes, our decoder framework produces meaningful information from an input RSI.

4 Experimental Results

In this section, we present details of the datasets used, implementation details, experimental results and comparative analysis to exhibit the performance of the

proposed approach.

4.1 Dataset Details

4.1.1 UCM-captions dataset

The UCM-Caption dataset [Yang Y, 2010] consists of 21 classes of land use images with 100 images in each class. Each image is of 256x256 size and was extracted from the United States Geological Survey (USGS) National Map. Every image in this dataset is described through five sentences. All the five sentences for each image are diversified but the difference between sentences of the same class is very small. There are 2100 images with 10500 captions respectively.

4.1.1 Sydney-captions dataset

Sydney-captions dataset [Zhang F, 2014] was collected from Google Earth. It has 613 images of airports, residential areas, rivers etc., and are categorised into seven classes. Similar to the UCM-caption dataset, each image is

described with five sentences and the entire dataset has 3065 descriptions for those 613 images.

4.2 Parameter settings and implementation details

The proposed deformable and structural representative network (SRN) are implemented using the Pytorch framework. The dimension of structural representative features, the embedding of the attention layer, hidden LSTM, and feature maps are set to dimension 512. In our work, we implement ADAM optimizer with 0.0001 and 0.0003 learning rates for the visual encoder and caption decoder. We fixed batch size to 32 throughout the implementation. This model keeps on learning until the accuracy of the model remains the same and also completes 15 epochs on the validation set. At last, the decay rate is employed, when the model does not progress for 6 epochs. Further, the ResNet-101 [He K, 2016] backbone network is pre-trained on Imagenet and is used at the visual encoder. Initially, we obtain the spatial features from Conv3 and semantic features from Conv5 of backbone with resolutions of 32×32×512 and 8 × 8 × 2048, respectively. Then, the spatially transformed features are captured from a deformable network employed on top of initial layers with two deformable and standard convolutions. For standard convolutions, We set bottleneck layer filters with 512 channels and 3×3 filters for the deformable convolutions. Later, the SRN is incorporated on the Conv5 feature map of backbone with various dilated convolutions. SRN consists of 1×1 conv and 3×3 conv layers with different dilation rates ie., 2, 4, and 6. A single conv layer of 1×1 and average pooling is included in the network along with dilated convolutions to obtain features from various fields of views. Before fusing, the spatially transformed and semantic features obtained from two networks, the channel count was reduced to 512.

4.3 Quantitative Results

Our approach is evaluated on two datasets namely Sydney-captions and UCM-captions. Our method is compared with various well-known methods such as VLAD + RNN [Jégou H, 2011], GloVe [Pennington J, 2014], Hard-attention, Soft-attention [Lu X, 2017], mRNN [Gu J, 2018], ConvCap [Aneja J, 2018], mGRU-embedword [Lu X, 2019], CSMLF [Wang B, 2019], SAA [Lu X, 2019], and RTRMN[Wang B, 2020]. In our work, we split data as 80% for training, 10% for validation, and the rest for testing.

Evaluation Metrics: For evaluating the accuracy of the produced captions, four different metrics were used such as ROUGE-L [Lin CY, 2004], CIDEr-D [Vedantam R, 2015], METEOR [Banerjee S, 2005], and BLEU [Papineni K, 2002]. These metrics are widely used in almost all image captioning tasks.

4.3.1 BLEU

The co-occurrences among the generated and the ground truth captions is measured using the BLEU [Papineni K, 2002] that considers a sequence of n number of ordered words. The BLEU-n (n = {1, 2, 3, 4}) is computed as the ratio of the n-grams that are matched with the total number of n-grams in the evaluated caption to the total n-grams. This scores in the range of 0.0 and 1.0

4.3.2 ROUGE-L

ROUGE-L [Lin CY, 2004] is an updated form of ROUGE, which computes an F-measure that uses the longest common subsequence (LCS) between the generated and the ground-truth captions with a recall bias. This scores in the range of 0.0 and 1.0

4.3.3 CIDEr-D

CIDEr-D [Vedantam R, 2015] is another version of CIDEr, where initially, the caption is converted into the term frequency inverse document frequency (TF-IDF) vector [Robertson S, 2004] and then the reference caption is produced using the cosine similarity and finally the caption is generated by the model. CIDEr-D regularises n-grams repetition if they occur beyond the number of times in the reference sentence. Higher score indicates a higher accuracy. The score of this metric is between 0.0 and 1.0.

The evaluation scores presented in Tables 1 and 2 follow similar experimental procedures i.e. 80% for training, 10% for testing, & 10% for validation, making the evaluation fair. It is observed that our approach performs well when compared to the existing methods due to the addition of rich semantics and textual information.

Table 1: Evaluation scores (%) on the SYDNEY - CAPTIONS dataset.

Method	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-L	CIDEr-D
mGRU +embedword [Lu X, 2019]	68.85	60.03	51.81	44.29	57.47	168.94
VLAD +LSTM [Lu X, 2017]	49.13	34.12	27.60	23.14	42.01	91.64
VLAD+RNN [Lu X, 2017]	56.58	45.14	38.07	32.79	52.71	93.72
ConvCap [Aneja J, 2018]	74.72	65.12	57.25	50.12	66.74	214.84
Soft-attention [Lu X, 2017]	73.22	66.74	66.23	58.20	71.27	249.93
Hard-attention [Lu X, 2017]	75.91	66.10	58.89	52.58	71.89	218.19
CSMLF [Wang B,2019]	59.98	45.83	38.69	34.33	50.18	75.55

SAA [Lu X, 2019]	68.82	60.73	52.94	43.89	58.20	175.52
Ours	**77.85**	**68.01**	**68.51**	**60.40**	**73.64**	**228.56**

Table 2: Evaluation scores (%) on the UCM-CAPTIONS dataset [Yang Y, 2010] .

Method	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-L	CIDEr-D
mGRU+ embedword [Lu X, 2019]	75.74	69.83	64.51	59.98	66.74	279.2
VLAD+ LSTM [Lu X, 2017]	70.16	60.85	54.96	50.30	65.20	231.3
VLAD+RNN [Lu X, 2017]	63.11	51.93	46.06	42.09	58.78	200.6
ConvCap [Aneja J, 2018]	70.34	56.47	46.24	38.57	59.62	190.1
Soft-attention [Lu X, 2017]	74.54	65.45	58.55	52.50	72.37	261.24
Hard-attention [Lu X, 2017]	81.57	75.12	67.02	61.82	76.98	299.47
CSMLF[Wang B,2019]	36.71	14.85	7.63	5.05	29.86	13.51
SAA [Lu X, 2019]	79.62	74.01	69.09	64.77	69.42	294.51
Ours	**82.92**	**76.95**	**72.17**	**67.75**	**78.49**	**315.15**

4.4 Qualitative Results

The qualitative results obtained due to the proposed approach is shown in Figures 3 and 4, where each word generated from the decoder has appropriate attention mask for the given RSI. As observed in Figures 3 and 4, the produced words have appropriate mapping with corresponding image parts through an attention map. Our approach generates highly descriptive, accurate, very precise words by encoding spatial and structural representative features of RSI. The visualisations of the captioning results are shown in Figures 3 and 4 respectively.

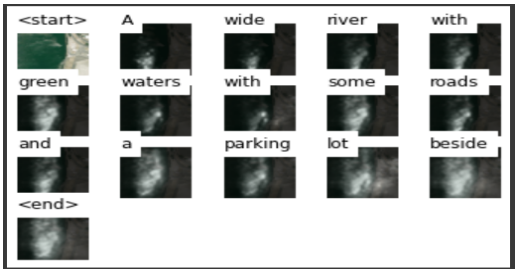


Figure. 3. Generated caption with attention mask on the UCM-Caption dataset.

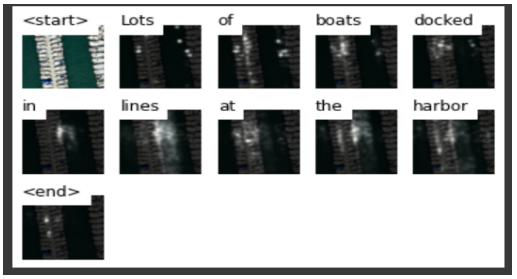


Figure. 4. Generated caption with attention mask on the Sydney-Captions dataset.

5 Conclusions

For holistic representation of an RSI, most image captioning methods utilise image-level features or visual entities. However, these approaches cannot incorporate multi-scale structural information and spatial features of small entities. To overcome this problem, a novel deformable and structural representative network (SRN) is proposed for remote sensing image captioning. Particularly, the semantically spatial features are obtained from the backbone network. Subsequently, we have developed a deformable network on the initial layers and

SRN on the last layers of the CNN to obtain spatially transformed information and structural representations of an RSI. Finally, a multi-stage caption decoder is utilised to produce meaningful captions. In our approach, a stack of LSTMs in the decoder helps to deal with the vanishing gradient problem and also includes mid-path monitoring. Our approach performed better than the well-known RSIC methods.

References

X. Zhang, X. Li, J. An, L. Gao, B. Hou, and C. Li, “Natural language description of remote sensing images based on deep learning,” in 2017 *IEEE International Geoscience and Remote Sensing Symposium*, IGARSS 2017, Fort Worth, TX, USA, July 23-28, 2017, 2017, Conference Proceedings, pp. 4798–4801.

X. Lu, B. Wang, and X. Zheng, “Sound Active Attention Framework for Remote Sensing Image Captioning,” *IEEE Trans. Geosci. Remote. Sens.*, vol. 58, no. 3, pp. 1985–2000, 2020. [Online]. Available: https://doi.org/10.1109/TGRS.2019.2951636.

R. Zhao, Z. Shi, and Z. Zou, “High-resolution remote sensing image captioning based on structured attention,” *IEEE Transactions on Geo-science and Remote Sensing*, pp. 1–14, 2021.

Q. Wang, W. Huang, X. Zhang, and X. Li, “Word-sentence framework for remote sensing image captioning,” *IEEE Transactions on Geoscience and Remote Sensing*, pp. 1–12, 2020.

X. Ma, R. Zhao, and Z. Shi, “Multiscale methods for optical remote-sensing image captioning,” *IEEE Geoscience and Remote Sensing Letters*, pp. 1–5, 2020.

Z. Shi and Z. Zou, “Can a machine generate human-like language descriptions for a remote sensing image?” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 6, pp. 3623–3634, 2017.

X. Lu, B. Wang, X. Zheng, and X. Li, “Exploring models and data for remote sensing image caption generation,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 4, pp. 2183–2195, 2018.

X. Zhang, X. Wang, X. Tang, H. Zhou, and C. Li, “Description generation for remote sensing images using attribute attention mechanism,” *Remote Sensing*, vol. 11, no. 6, p. 612, 2019.

Z. Zhang, W. Diao, W. Zhang, M. Yan, X. Gao, and X. Sun, “Lam:Remote sensing image captioning with label-attention mechanism,” *Remote Sensing*, vol. 11, no. 20, p. 2349, 2019.

G. Sumbul, S. Nayak, and B. Demir, “Sd-rsic: Summarization-driven deep remote sensing image captioning,” *IEEE Transactions on Geo-science and Remote Sensing*, pp. 1–13, 2020.

Ma, Chao, Jia-Bin Huang, Xiaokang Yang, and Ming-Hsuan Yang.,“Robust visual tracking via hierarchical convolutional features,” *arXiv preprint* arXiv:1707.03816, 2017.

Jaderberg, Max, Karen Simonyan, and Andrew Zisserman., “Spatial transformer networks,” In Advances in *neural details processing systems*, pp. 2017-2025, 2015.

Si, Haiyang, Zhiqiang Zhang, Feifan Lv, Gang Yu, and Feng Lu., “Real-Time Semantic Segmentation via Multiply Spatial Fusion Network,” *arXiv preprint* arXiv:1911.07217, 2019.

Chen, Liang-Chieh, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam., “Encoder-decoder with atrous separable convolution for semantic image segmentation,” in *Proceedings of the European conference on computer vision (ECCV)*, pp. 801-818, 2018.

Wang, B., Zheng, X., Qu, B. and Lu, X., 2020. Retrieval topic recurrent memory network for remote sensing image captioning. *IEEE Journal of Selected Topics in App. Earth Obs. and Remote Sensing*, *13*, pp.256-270.

Gu, Jiuxiang, Jianfei Cai, Gang Wang, and Tsuhan Chen., “Stack-captioning: Coarse-to-fine learning for image captioning,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1. 2018.

K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. Zemel, and Y. Bengio, “Show, attend and tell: Neural image caption generation with visual attention,” in *Int. Conf. on machine learning*, 2015, pp. 2048–2057.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and ZbigniewWojna., “Rethinking the inception architecture for computer vision,” in Proceedings of the *IEEE conference on computer vision and pattern recognition*, pp. 2818–2826, 2016.

Wei, Haiyang, Zhixin Li, Canlong Zhang, Tao Zhou, and Yu Quan., “Image captioning based on sentence-level and word-level attention,” in 2019 *Int. Joint Conference on Neural Networks (IJCNN)*, pp.1-8. IEEE, 2019.

He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun., “Deep residual learning for image recognition,” in Proceedings of the *IEEE conference on computer vision and pattern recognition*, pp. 770-778.2016.

Vinyals, Oriol, Alexander Toshev, Samy Bengio, and Dumitru Erhan., “Show and tell: A neural image caption generator,” in Proceedings of the *IEEE conference on computer vision and pattern recognition*, pp. 3156-3164. 2015.

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

Chen, Long, Hanwang Zhang, Jun Xiao, Liqiang Nie, Jian Shao, Wei Liu, and Tat-Seng Chua., “Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning,” in Proceedings of the *IEEE conference on computer vision and pattern recognition*, pp. 5659-5667. 2017.

J. Aneja, A. Deshpande, and A. G. Schwing, “Convolutional image captioning,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit*., Jun. 2018, pp. 5561–5570.

H. Jegou, F. Perronnin, M. Douze, J. Sánchez, P. Perez, and C. Schmid, “Aggregating local image descriptors into compact codes,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 9, pp. 1704–1716, Sep. 2012.

J. Pennington, R. Socher, and C. Manning, “Glove: Global vectors for word representation,” in *Proc. Conf. Empirical Methods Natural Lang. Process.* (EMNLP), 2014, pp. 1532–1543.

B. Wang, X. Lu, X. Zheng, and X. Li, “Semantic descriptions of high-resolution remote sensing images,” *IEEE Geosci. Remote Sens. Lett*., vol. 16, no. 8, pp. 1274–1278, Aug. 2019.

K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, “BLEU: A method for automatic evaluation of machine translation,” in Proc. 40th Annu. *Meeting Assoc. Comput. Linguistics (ACL)*, 2001, pp. 311–318.

C.-Y. Lin, “Rouge: A package for automatic evaluation of summaries,” in Text Summarization Branches Out. Barcelona, Spain: *Association forComputational Linguistics*, 2004, pp. 74–81.

A. Lavie and A. Agarwal, “Meteor: An automatic metric for MT evaluation with high levels of correlation with human judgments,” in *Proc. 2nd Workshop Stat. Mach. Transl. StatMT*, 2007, pp. 65–72.

R. Vedantam, C. L. Zitnick, and D. Parikh, “CIDEr: Consensus-based image description evaluation,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, June. 2015, pp. 4566–4575.

Yi-Nung Chung, Tun-Chang Lu, Ming-Tsung Yeh, Yu-Xian Huang, and Chun-Yi Wu, "Applying the Video Summarization Algorithm to Surveillance Systems," *Journal of Image and Graphics*, Vol. 3, No. 1, pp. 20-24, June 2015. doi: 10.18178/joig.3.1.20-24

S. Robertson, “Understanding inverse document frequency: On theoretical arguments for IDF,” *J. Document*., vol. 60, no. 5, pp. 503–520, Oct. 2004.

Dai, Jifeng, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei., “Deformable convolutional networks,” in Proceedings of the IEEE international conference on computer vision, pp. 764-773. 2017.

Lu, Yu, Muyan Feng, Ming Wu, and Chuang Zhang., “C-DLinkNet: considering multi-level semantic features for human parsing,” arXiv preprint arXiv:2001.11690, 2020.

Jaderberg, Max, Karen Simonyan, and Andrew Zisserman., “Spatial transformer networks,” In Advances in neural information processing systems, pp. 2017-2025, 2015.

Y. Yang and S. D. J. a. i. g. i. s. Newsam, “Bag-of-visual-words and spatial extensions for land-use classification,” in 18th ACM SIGSPATIAL International Symposium on Advances in Geographic Information Systems, ACM-GIS 2010, November 3-5, 2010, San Jose, CA, USA, Proceedings, 2010, Journal Article, pp. 270–279.

F. Zhang, B. Du, and L. Zhang, “Saliency-guided unsupervised feature learning for scene classification,” IEEE Transactions on Geoscience and Remote Sensing, vol. 53, no. 4, pp. 2175–2184, 2014.

Ramos, Rita, and Bruno Martins. "Using Neural Encoder-Decoder Models With Continuous Outputs for Remote Sensing Image Captioning." IEEE Access 10 (2022): 24852-24863.

Wang, Qi, Wei Huang, Xueting Zhang, and Xuelong Li. "Word–sentence framework for remote sensing image captioning." IEEE Transactions on Geoscience and Remote Sensing 59, no. 12 (2020): 10532-10543.

Banerjee, Satanjeev, and Alon Lavie. "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments." *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*. 2005.