Structural representation network for remote sensing image captioning

Jaya Sharma, Peketi Divya, Yenduri Sravani, Krishna Mohan C.
Department of Computer Science and Engineering, Indian Institute of Technology Hyderabad, Hyderabad, India.
{cs18m19p100002@iith.ac.in, ai21resch01001@iith.ac.in, cs18resch02001@iith.ac.in, ckm@cse.iith.ac.in}

Abstract

Current encoder-decoder methods for remote sensing image captioning ignore the fine-grained structural representation of objects due to the lack of prominent encoding frameworks. Hence, in this paper, we propose a novel structural representation network (SRN) to capture the fine-grained structures of remote sensing images (RSI) for generating semantically significant captions. Initially, we employ SRN on top of the final layers of the convolutional neural network (CNN) to attain the features of RSI that are spatially transformed. Next, an attention mechanism is utilized on SRN to obtain dense features. Then, a multi-stage decoder is incorporated into the extracted dense features to produce fine-grained meaningful captions. The efficacy of the proposed approach is demonstrated on three RSI captioning datasets i.e., UCM-Captions, Sydney-Captions, and RSICD dataset.

Keywords: Attention mechanism, multi-stage LSTM, remote sensing image captioning, structural representation network.

1. INTRODUCTION

The fastest growth in technology on artificial satellites and launch vehicles makes it simple to recognize the lands and earth. Nevertheless, an enormous number of high-resolution remote sensing images (RSI) are generated. Therefore, it is challenging for users to access, store, and particularly observe relevant data from a large amount of information that addresses the client's issues.

Remote sensing image captioning (RSIC) is a task that emulates the text from the imaging modality and provides meaningful captions that describe the content of remote sensing images (RSI). It is a probabilistic approach conditioned on the image's attributes and visual features. The main goal of RSIC is that the generated captions need to handle the relationship between the scene and the object of an image. Thus, it became active research in fields such as image interpretation & understanding, text retrieval in an image, information generation [20], robotic vision, and content search [21], [22], [23], [24].

Several remote sensing image captioning methods in the literature explore encoder-decoder frameworks to generate meaningful captions. These methods produce sentences by arranging all the words predicted by the recurrent neural networks (RNNs) in a sequence, depending on the remote sensing image features captured by convolution neural networks (CNNs). Gu et al. [25] have proposed a coarse-to-fine multi-stage caption generation network consisting of multiple long short-term memory (LSTM) decoders to produce meaningful captions. Recently, the attention mechanism [26] has been applied to the encoder and decoder of the models to focus on important parts of an image. The transformer-based architecture is used in [27] to enhance the encoding and caption generation tasks. Transformer [31] has been very popular in processing sequential tasks. This method not only be competent at sequence tasks, such as, machine translation [32], [33], question-answer [34], and text classification [35], but also applied to many vision tasks, classification [36], and object detection [37], [38], [39].

The encoder-decoder setup utilizes CNNs to capture image visual data, and RNNs are employed to generate sentences. In the wide range of recent works, captions are generated based on the semantic information obtained at the last layer of CNN's [28], [29], [39]. Nevertheless, these methods are not efficient in generating meaningful captions due to the lack of structural representation of an RSI. The primary challenges in current approaches are: (i) inefficient spatial...
and structural representations of RSI and (ii) the use of a single-stage caption decoder that fails to produce meaningful captions.

To overcome the above challenges, a structural representation network (SRN) is proposed. The basic overview of SRN is shown in Figure 1. Basically, the last layers of CNNs provide semantic information but do not capture the boundary information of an object due to repetitive operations such as pooling & striding. Thus, we build an SRN on top of the final layers of CNN to retrieve a structural representation of an RSI by applying dilated convolutions parallelly at various rates in different field-of-views. To obtain more dense features, an attention mechanism is applied. Lastly, spatial features are fed to the multi-stage decoder network to obtain meaningful descriptions of an RSI.

The main contributions of our work are summarized as follows:

- We propose a structural representative network (SRN) to effectively capture the fine-grained structures of remote sensing images for generating captions.
- The proposed encoder-decoder framework incorporates spatial invariant features by exploiting structural representative networks.
- Our proposed model produces meaningful captions by capturing efficient representations of an image and its efficiency is demonstrated on three remote sensing image captioning datasets, namely, UCM-Captions, Sydney-Captions, and RSICD.

The rest of the paper is arranged as follows. section 2 is an overview of existing literature works devoted to remote sensing image captioning and related frameworks. In Section 3, we describe the proposed methodology. In Section IV, we discussed datasets and implementation details. In Section V we showed quantitative and qualitative results and finally, the conclusion is provided in section VI.

2. RELATED WORKS

This section, reviews the state-of-the-art remote sensing image captioning frameworks and discusses the roles of different statistical learning algorithms in structure analysis and semantics extraction.

2.1 Image Captioning in Remote Sensing Images:

Initially, Zhang et al. [1] proposed a CNN-RNN-based approach for image captioning tasks. In this method, CNNs are utilized to extract features of the remote sensing images (RSI), and recurrent neural networks (RNN) are used in predicting words in a sequence for providing meaningful descriptions. There are still promising approaches for producing descriptions for remote sensing images (RSI) [2], [3], [4], [5]. Recently, Z. Shi et al. [6] proposed a template-based method that uses CNN and FCN [7] to capture the contents of an image. Later, the captured words are composed into complete sentences with stable templates. X. Lu et al. [8] introduced three RSI captioning datasets and carried out a sequence of evaluations on these datasets. An increase in performance has been observed due to the soft attention mechanism. Later, attribute attention models [9] are proposed where these models use the attribute data captured from the higher-level features of an image to reweight the features of an image for producing a large number of silent features. Thus the model can use this attribute data for improving the model's strength in generating sentences. Recently, a label-guided attention mechanism was proposed by Z. Zhang et al. [10] where the information of the label is used in the process of attention calculation to adjust the weights of the attention layer again. In this process, the image data is filtered without any relevance to the label of RSI. Thus, more meaningful sentences are generated by providing more salient features of an RSI at the decoding steps. Next on, in order to improve the salience features of an RSI, a new visual aligning loss was proposed in the training process by [11]. Later, a new RSI captioning approach was proposed by [12] for summarizing the ground truth captions. In this method, firstly, a standard CNN-RNN pipeline was adopted for predicting words at every step. Then it uses the summarization model for summarizing all five ground truth captions into one caption. Lastly, in order to attend to predicted words dynamically in the standard and summarization time steps, the model uses the adaptive weighting strategy.

2.2 Visual feature encoding methods
The rich high-level features are exploited in [13] using deep convolutional neural networks (CNN) to improve the robustness and accuracy of visual tracking. These rich features describe the image pyramidal representations at different abstraction stages. The coarse and semantic data are concatenated using the fusion method in [14] to obtain the spatial aware visual features of an image. Another fusion method named ExFuse was proposed for the segmentation task by [15] where they integrated semantic information into low-level features of an image, and high spatial resolution features into high-level features of an image. Spatial transformer networks (STNs) are proposed by Jaderberg et al. [16] to learn the spatially invariant features of an image. This model removes affine or perspective spatial transformation and also incorporates spatial manipulations in the data. Later, Yu and Koltun [17] came up with dilated convolutions in CNNs to introduce multi-scale contextual features of an image. These dilated convolutions capture in-depth contextual information by exponentially increasing the size of the receptive field. H Si et al. [18] proposed a multiply spatial fusion network (MSFNet) to extract spatial information and increase receptive field size. Finally, the spatial pyramid pooling module is introduced in [19] to probe dilated convolutions parallelly with multiple fields of view at various rates.

![Figure 1: Basic overview](image)

3. PROPOSED APPROACH

In this section, firstly we explain the classical encoder-decoder framework for generating the caption. In this encoder-decoder framework, a convolutional neural network (CNN) is employed at the encoder to encode the visual depiction of an image and at the decoder, long short term memory (LSTM) is used to decode the visual attributes and produce a sequence of words. Then, the proposed structural representative network is utilized for the task of image captioning.

Department of Computer Science and Engineering, Indian Institute of Technology Hyderabad, Hyderabad, India.

3.1. Caption generation using Encoder-Decoder framework

Given an input image $G$ and its corresponding caption $C = \{\omega_1, \omega_2, \ldots, \omega_N\}$ where, weight kernel matrix $\omega$ consists of $N$ number of words, the objective function $\alpha$ is maximized by the encoder-decoder model as $\sum \alpha$.
\[ \alpha^* = \arg \max \log p(C|G), \quad (1) \]

where \( \alpha \) are the model parameters, \( I \) is the input feature map. Then, the chain rule is used to define the log-likelihood of joint probabilities of all words as

\[ \sum_{\bullet \bullet = 1} \log p(C|G) = \log p(w_i|w_1, \ldots, w_{i-1}, G) \quad (2) \]

The parameters of the model are dropped here for convenience. Later, using the encoder-decoder framework, every conditional probability is designed as

\[ \log p(w_i|w_1, \ldots, w_{i-1}, G) = f(h_i, c_i), \quad (3) \]

where a nonlinear output function \( f \) produces the probability of every predicted word \( \omega_i \). The \( c_i \) and \( h_i \) are context vectors and hidden states of RNN at time interval \( i \). Then, a Long-short term memory (LSTM) network is used to produce a caption of a remote sensing image. In an LSTM network, the hidden state \( h_i \) is modeled as

\[ h_i = LSTM(a_i, h_{i-1}, c_{i-1}) \quad (4) \]

where the input vector is \( a_i \). Generally, the context vector is modeled by considering two options i.e., attention-based and vanilla encoder-decoder frameworks, and it serves as visual proof for producing the caption of an image.

In a nutshell, the vanilla encoder-decoder framework selects a context vector from the last fully connected layer of the convolutional network. Throughout the process of caption generation, the context vector will be similar and doesn’t count on the data acquired by the RNN decoder module. Although the context vector-based attention encoder-decoder framework relies upon both caption decoder and visual encoder modules, it focuses on essential regions of an image at every step of the hidden state of RNN. In this work, the attention encoder-decoder-based framework is adopted to produce the assisted contextual features for the task of image captioning.

Figure 2. The framework of the proposed Structural representative network (SRN).

3.2. Structural representative network (SRN) for image captioning
In this work, a structural representative network (SRN) is proposed for the task of caption generation to encode features of an input remote sensing image that are transformed spatially. The SRN presented in Figure 2 adapts an encoder-decoder framework with attention, where the visual encoder consists of the structural representative network (SRN) and recurrent criss-cross attention mechanism (RCCAM). And, the caption decoder consists of a Long-short term memory (LSTM) network. The structural representative network is applied above the semantic features that are extracted from the topmost layer of the backbone network to assimilate the multi-scale contextual information. Later, on the structural representative network, a criss-cross attention mechanism [40] is applied to attend to noticeable regions of visual scene content. Finally, the attention mechanism [41], a multi-stage LSTM network is utilized that produces the words which are directed by attentive contextual features. The visual encoder and caption decoder modules are described well in the below subsections.

3.3. Visual encoder

It is a feature encoding network where the multiscale scene contextual information and spatially transformed features of an image are encoded by processing the input remote sensing image through distinct network components namely, backbone network, structural representative network, feature fusion, and criss-cross attention mechanism.

3.3.1. Backbone network: To retrieve the visual depiction of an image, a pre-trained ResNet [42] is used as a backbone feature extraction network which has five layers, namely, Conv1, Conv2, Conv3, Conv4, and Conv5, where each of the Conv layers is composed with a various number of bottle-neck layers. Generally, abundant spatial features of small entities are held by the initial layers of the backbone network but it fails to assimilate semantic information of an image. Although the final layer consists of potential semantic information, it lacks spatial resolutions of entities. In Figure 2, the framework for the task of image captioning is shown, where the fully connected layers of the ResNet backbone network are removed and the spatial and semantic features of Conv layers are used. Moreover, the semantic features are achieved from the Conv5 layer of ResNet with $8 \times 8$ resolution and 2048 channels. Additionally, a structural representative network is employed above the Conv5 layer to assimilate multi-scale contextual information of an image.

3.3.2. Multi-scale structural representative network: Generally, by using the semantic information of an image that is retrieved from the final convolutional layer of the backbone network, mostly all the image captioning works [43], [44], [45], [46], produces the caption. Even though the final CNN layer consists of rich semantic details, the subtle information of object boundaries reduces due to the stridden convolutional and multiple pooling applications. A structural representative network is employed above the Conv5 layer of the backbone network to address this issue. Particularly, in the structural representative network, various parallel dilated convolutions [47] are employed on the output feature map of Conv5 with distinct scales for capturing scene contextual details. The structural representative network helps us to manage the receptive field size of input feature maps and also to expand the filter’s field of view. Without raising the learnable parameters and computation time, the greater contextual details of an image are employed in the network. Also, to sample the input receptive fields at multiple field-of-views and multiple rates, the parallel dilated convolutions support segmenting the objects at multiple scales and condition every layer.

A two-dimensional signal is given for each location $l$ on the output feature map $m$ and weight kernel matrix $ω$, the dilated convolutions are used on input feature map $k$ as

$$ m[l] = k[l + d \cdot x]ω[x] $$

where $d$ denotes the rate of dilation at which the input feature map is sampled. $d = 1$ is the special case of dilated convolutions that denotes the standard convolution. The filter’s field-of-view alters adaptively as the dilation rate gets changed.

3.3.3. Recurrent criss-cross attention: Even though the CNN models are considered an unusually capable class of models, due to the limited geometric structure and providing short-range contextual details, they are internally bounded within local receptive fields. To focus on this issue, a structural representative network with dilated convolutions at
multiple scales is proposed. However, scene contextual details are collected by our structural representative network at numerous scales but lack dense contextual features [40]. For the attention method, numerous works are utilized [48], [49], to produce dense contextual details by collecting the contextual details at every point via produced attention maps. To produce the attention-guided contextual features, the output of the structural representative network is fed into a criss-cross attention module [40]. When compared to conventional attention systems [45], [50], the criss-cross attention system is a memory-friendly system and substantially decreases FLOPs. Particularly, in a criss-cross fashion, by almost connecting one pixel to the other pixels it restores the dense non-local attention blocks with the sparse attention. Moreover, recurrent operation of the criss-cross module is used to capture the full image dependencies. The local context information is collected by the first attention module in horizontal and vertical directions. Then, the additional details from more augmented pixels are collected by the recurrent attention module and the full image dependencies are captured.

Given, that \( I \) is an input feature map and \( M \) is obtained by using a convolutional layer of a feature map. Then, to get a new feature map \( M' \) that extracts the contextual details in the vertical and horizontal directions (criss-cross path), the feature map \( I \) is fed to the criss-cross attention module. To achieve a dense contextual feature, the above procedure is performed recurrently and a new feature map \( M' \) is obtained. In the criss-cross attention system, given, a feature map \( M \in \mathbb{R}^{C \times W \times H} \), two feature maps \( J \) and \( L \) are generated using \( I \times I \) convolutions, where \( \{ J, L \} \in \mathbb{R}^{C \times W \times H} \). Further, a feature vector \( Q_{o} \in \mathbb{R}^{e} \) and \( U_{o} \in \mathbb{R}^{(H+W-1) \times C} \) is obtained from each position \( o \) in the spatial domain of \( J \) and \( L \), respectively. To produce attention map \( A \), first, the affinity operation is performed as

\[
r_{v, o} = J_{v, o} \cdot U_{v, o}, \tag{6}
\]

where, \( r_{v, o} \in \mathbb{R} \) is the correlation between features \( J_{v, o} \) and \( U_{v, o} \) (\( e \) is \( e^{th} \) element of \( U_{v, o} \)). Then, softmax is applied on layer \( R \).

Additionally, another feature map \( N \in \mathbb{R}^{C \times W \times H} \) is produced by employing a convolutional layer on \( M \). Further, a feature map \( U_{v} \in \mathbb{R}^{(H+W-1) \times C} \) is obtained by collecting the features in the same column or row at each position \( U_{v} \). Then, the contextual details are aggregated to produce an output feature vector \( M' \) as

\[
\sum_{o} Y_{o} M'_{v, o} = R_{v, o} + M_{o}, \tag{7}
\]

where, the output feature map \( M' \) contains meaningful contextual details. Further, the same criss-cross attention mechanism is repeated with the \( M' \) to obtain rich contextual information \( (M'') \).

### 3.4. Caption decoding module

Generally, most image captioning frameworks utilize a one-stage caption decoder module to produce a caption of an image. However, due to the shortage of transitional supervision, they are unsuccessful in producing rich fine-grained information. To diminish this issue and efficiently employ our visual encoder depiction, a coarse-to-fine multistage caption decoder framework [41] is used for caption generation. Usually, it is outfitted with numerous decoders where every decoder element performs on the output of the prior stage and generates precise image information successively. Additionally, it focuses on the issue of vanishing gradients that arose due to multi-stage models and emphasizes intermediate supervision through reinforcement learning. In Figure 2, the architecture of the coarse-to-fine multi-stage caption decoding module is illustrated. Along with attention modules, it contains three stacked long-short term memory (LSTM) networks, as shown in the Figure. Particularly, at stage one of the LSTM decoder, coarse-grained image information is produced, and the succeeding LSTM decoding network generates the fine-grained information. At every stage of the model, attention weights and preceding stage hidden vectors are given as input to produce more precise captions.

Firstly, the coarse decoder LSTM network \( (LSTM_c) \) is learned on accomplishing encoded details of an image by utilizing the visual encoder. At every time interval \( i \), the details of preceding words, the visual depiction of an image, and the previous hidden states of the LSTM network are taken by LSTM, to produce the caption as
\[ c_i^0, h_i^0 = \text{LSTM}(h_{i-1}^0, x_{i-1}^0, w_{i-1}), \]
\[ x_i^0 = f(Z; h_{i-1}^0), \]
\[ (8) \]

where the hidden states are defined as \( h_{i-1}^0 \) and \( h_{i-1}^0 \), the cell state is \( c_i^0 \), the preceding word is \( w_{i-1} \). \( x \) denotes \( x = 0 \) for \( \text{LSTM} \), and \( x \geq 1 \) for fine decoders \( \text{LSTM}_F \), the total amount of fine stages are indicated by \( N_f \), and the mean pool of visual encoder features is denoted by \( f(Z) \). Further, utilizing the attention weights \( \alpha^{i-1} \), fine stage decoders, visual information, and preceding words are precisely captioned as
\[ c_i^0, h_i^0 = \text{LSTM}_F(h_{i-1}^0, x_{i-1}^0, w_{i-1}), \]
\[ x_i^0 = g(Z, \alpha^{i-1}, h_{i-1}^0, h_{i-1}^0), \]
\[ (12) \]

where \( g(.) \) denotes the function of spatial attention that generates attention guided visual information. On achieving attentive features, the coarse-to-fine LSTM network generates the fine-grained description of an image.

### 4. EXPERIMENTAL RESULTS

In this section, firstly, the public RSIC datasets and the common evaluations are described. Then, the implementation details are presented. Finally, the analysis of experimental results is given.

#### 4.1. Datasets

Three RSIC datasets, namely, UCM-Captions, Sydney-Captions, and RSICD, are used for our experiments. For the task of scene classification, all the images used in these datasets are shot and prepared to utilize the remote satellites. Later, with the generated sentences, X. Lu et al. [51] accomplished these three RSIC datasets and published them as RSIC datasets. Each dataset is split into three parts, for training 80% of image-caption combinations are used, for validating 10% of data is utilized, and 10% is used for testing.

**UCM-Captions:** UCM-Captions is an RSIC dataset that has been revised by the UC Merced Land Use Dataset [52], the National Map of the United States Geological Survey (USGS) [53] shot it, and used for the task of scene classification. It has 21 scene classes with 100 images in each class and the size of an image is 256*256 pixels. For each image, five different meaningful sentences are generated and this dataset is named as UCM-Captions.

**Sydney-Captions:** The Sydney-Captions [54] is created with 7 classes, 500 * 500 pixels of image size, and 613 images in total. Moreover, X. Lu et al. [51] constructed the Sydney-Captions dataset by accomplishing five different meaningful sentences for every image.

**RSICD:** Presently, RSICD is the largest among all three datasets when it is considered for the remote sensing image caption dataset. The images have been shot by airplanes and satellites simultaneously. RSICD has a total of 10,921 images of size 224 * 224 pixels [55] with 5 annotations for each image. While preparing RSICD, to improve the accuracy and diversity of image annotations, various guidelines for annotation are made and followed in representing the content of images [51]. Thus, RSICD is accurate in the information and complete in content.

**Evaluation Metrics:** For evaluating the accuracy of the produced captions, four different metrics were used including the bilingual evaluation understudy ROUGE-L [75], CIDEr-D [77], METEOR [76], and (BLEU) [74], which are all widely used in recent image captioning literature.

**BLEU:** The co-occurrences between the generated caption and the ground truth is measured by the BLEU [74] using n-grams (which is a set of n ordered words). The key of the BLEU-n \( (n = \{1, 2, 3, 4\}) \) is the n-gram precision—the proportion of the matched n-grams out of the total number of n-grams in the evaluated caption.

**ROUGE-L:** ROUGE-L [75] is a modified version of ROUGE, which computes an F-measure with a recall bias using the
The VGG-16 a GRU the decoder.

For any of the above four metrics, a higher score indicates a higher accuracy. The scores of BLEU, ROUGE-L are between 0 and 1.0. The score of CIDEr-D is between 0 and 10.0.

4.2. Implementation details

We mainly executed the proposed structural representative network (SRN) by using the Pytorch [56] framework. 512 dimensions are fixed in the presented SRN model for the contextual embedding of the features, the embedding of the attention layer, and hidden LSTM and context vectors. Moreover, the ADAM optimizer [57] is utilized with a 0.0001 learning rate for the visual encoder and 0.0003 for the caption decoder. All through the network, the batch size is set to 32 and learned until the precision of the model don't get modified for 15 epochs on the validation set. At last, we utilize the decay rate when the model doesn't progress for 6 epochs.

We use the ResNet-101 [42] network as a backbone network for the visual encoder that has been pre-trained on Imagenet [58]. First, we collect the spatial semantic Conv5 attributes from the ResNet backbone network of size $8 \times 8 \times 2048$, respectively. For standard convolutions, a $1 \times 1$ filter with 512 channels is used. Then, the structural representation network is incorporated with various aligned dilated convolutions on the Conv5 attributes of the ResNet backbone network. At first, $1 \times 1$ conv and $3 \times 3$ conv is employed with dilation rates of 2, 4, and 6. Along with the dilated convolutions, a single conv layer of $1\times1$ channel and normal pooling is computed to assimilate attributes from different field-of-views. The conv layers of the complete structural representative network are set to 256 channels and all 5 layers are integrated before feeding it to the criss-cross attention module. Before feeding contextual attributes to the criss-cross attention module, the channel amounts are reduced to 512 and maintain the same amount of channels in the process.

4.3. Quantitative results

4.3.1. Comparison With Other Methods: Our method is evaluated on three data sets and has been compared with a variety of image captioning methods. The comparison methods include the VLAD + LSTM [67], mRNN [68], VLAD + RNN [67], mLSTM [68], mGRU-embedword [70], mGRU [78], ConvCap [63], Hard-attention [67], Soft-attention [67], CSMLF [69], SAA [70] and RTRMN [71]. Most of them, among these methods (ConvCap and except mGRU) are initially designed for the remote sensing image captioning task. However, their key ideas are mainly acquired from the natural image captioning [73], [72]. The description of these models are as follows.

**VLAD + RNN**: “VLAD” [64] handcrafted feature descriptors are used as its encoder by VLAD + RNN [67] to compute image representations and a naive RNN is used as its decoder to produce captions.

**VLAD + LSTM**: VLAD is also used by VLAD + LSTM [67] to compute the image features, but the difference here is that LSTM is used as its decoder.

**mGRU, mRNN, and mLSTM**: VGG-16 [65] is used by all these three methods [68], [78] as their encoders but RNNs used (naive RNN, LSTM, and GRU) as their decoders are different.

**mGRU-Embedword**: Same as the mGRU [78], the VGG-16 is also used by the mGRU-embedword [70] as its encoder and the GRU as its decoder. The difference here is that mGRU-embed word uses a pre-trained global vector, named as GloVe [66], to embed words.

**ConvCap**: The VGG-16 is used by the ConvCap [63] as its encoder and computes the attention weights based on the...
activations of the final convolutional layer. Rather than utilizing the RNN-based decoder, this technique produces captions by utilizing a CNN-based decoder [63].

**Soft-Attention and Hard-Attention:** Soft-attention [67] and Hard-attention [67] are two techniques involving VGG-16 as the encoder and LSTM as the decoder. The decoders are built based on soft attention and hard attention components [72], separately.

**CSMLF:** CSMLF [69] is a recovery based strategy that utilizes dormant semantic embedding to quantify the closeness between the image representation and the sentence representation in a typical semantic space.

**RTRMN:** RTRMN [71] utilizes Resnet-101 as its encoder and afterward utilizes the topic extractor to extract topic details. A retrieval topic recurrent memory network is utilized to produce captions in view of the subject words. "RTRMN (semantic)" and "RTRMN (statistical)" are two variations of the RTRMN, which depend on semantic topics repository and statistical topics repository, separately.

**SAA:** SAA [70] acquaints a SAA structure with join the sound data during the generating of captions. SAA involves the VGG-16 and sound GRUs as its encoder and involves other GRUs as its decoder.

**Baseline:** We first eliminate the proposed structured attention module of our technique and replace it with a standard soft-attention module [72] while keeping different designs unaltered as the baseline technique. Tables 1–3 shows the precision of our strategy and the above examination ones on the three different datasets.

All comparison strategies follow similar fixed segments of information (80% for training, 10% for validation, and 10% for test), which makes the examination fair. In these tables, the best scores are marked as bold. For the comparison techniques, the measurement scores are taken from the articles that proposed them. Since Qu et al. [68] didn't report the ROUGE-L scores on UCM-Captions and Sydney-Captions datasets these numbers are missing in Tables 1 and 2. We can see our strategy accomplishes the best accuracy in the majority of the entries. For instance, on the RSICD dataset, our baseline method (Resnet50 + LSTM + delicate consideration), which additionally applies beam search with a similar beam size during the inference stage, is now better compared to most of the other methods, as shown in Table 3. At the point when we incorporate the structured attention, we further work on our baseline by 2.97% on BLEU-4, 2.56% on METEOR, 1.30% on ROUGE-L, and 15.51% on CIDEr-D. While the baseline and proposed structured attention technique both uses ResNet-50 as the encoder and LSTM as the decoder, the proposed approach generally accomplishes a score higher than the baseline, and that implies that the accomplished improvement is because of the proposed structured attention strategy.

**TABLE 1: EVALUATION SCORES (%) OF DIFFERENT METHODS ON THE SYDNEY - CAPTIONS DATA SET**

<table>
<thead>
<tr>
<th>Method</th>
<th>BLEU-1</th>
<th>BLEU-2</th>
<th>BLEU-3</th>
<th>BLEU-4</th>
<th>ROUGE-L</th>
<th>CIDEr-D</th>
</tr>
</thead>
<tbody>
<tr>
<td>VLAD+RNN [64]</td>
<td>56.58</td>
<td>45.14</td>
<td>38.07</td>
<td>32.79</td>
<td>52.71</td>
<td>93.72</td>
</tr>
<tr>
<td>VLAD+LSTM [67]</td>
<td>49.13</td>
<td>34.12</td>
<td>27.60</td>
<td>23.14</td>
<td>42.01</td>
<td>91.64</td>
</tr>
<tr>
<td>mRNN [67], [68]</td>
<td>51.30</td>
<td>37.50</td>
<td>20.40</td>
<td>19.30</td>
<td>-</td>
<td>161.00</td>
</tr>
<tr>
<td>mLSTM</td>
<td>54.60</td>
<td>39.50</td>
<td>22.30</td>
<td>21.20</td>
<td>-</td>
<td>186.00</td>
</tr>
<tr>
<td>Method</td>
<td>BLEU-1</td>
<td>BLEU-2</td>
<td>BLEU-3</td>
<td>BLEU-4</td>
<td>ROUGE-L</td>
<td>CIDEr-D</td>
</tr>
<tr>
<td>------------------------------------</td>
<td>--------</td>
<td>--------</td>
<td>--------</td>
<td>--------</td>
<td>---------</td>
<td>---------</td>
</tr>
<tr>
<td>mGRU</td>
<td>69.64</td>
<td>60.92</td>
<td>52.39</td>
<td>44.21</td>
<td>59.17</td>
<td>171.55</td>
</tr>
<tr>
<td>mGRU+embedword [78]</td>
<td>68.85</td>
<td>60.03</td>
<td>51.81</td>
<td>44.29</td>
<td>57.47</td>
<td>168.94</td>
</tr>
<tr>
<td>ConvCap [63]</td>
<td>74.72</td>
<td>65.12</td>
<td>57.25</td>
<td>50.12</td>
<td>66.74</td>
<td>214.84</td>
</tr>
<tr>
<td>Soft-attention [67]</td>
<td>73.22</td>
<td>66.74</td>
<td>66.23</td>
<td>58.20</td>
<td>71.27</td>
<td>249.93</td>
</tr>
<tr>
<td>Hard-attention [67]</td>
<td>75.91</td>
<td>66.10</td>
<td>58.89</td>
<td>52.58</td>
<td>71.89</td>
<td>218.19</td>
</tr>
<tr>
<td>CSMLF [69]</td>
<td>59.98</td>
<td>45.83</td>
<td>38.69</td>
<td>34.33</td>
<td>50.18</td>
<td>75.55</td>
</tr>
<tr>
<td>SAA [70]</td>
<td>68.82</td>
<td>60.73</td>
<td>52.94</td>
<td>43.89</td>
<td>58.20</td>
<td>175.52</td>
</tr>
<tr>
<td>Baseline [72]</td>
<td>73.05</td>
<td>64.37</td>
<td>56.67</td>
<td>52.80</td>
<td>69.79</td>
<td>215.21</td>
</tr>
<tr>
<td>Structural representative network (ours)</td>
<td>73.28</td>
<td>65.02</td>
<td>57.90</td>
<td>51.84</td>
<td>65.37</td>
<td>197.40</td>
</tr>
</tbody>
</table>

* The “_” means that the scores are not reported in the reference papers.

TABLE 2: EVALUATION SCORES (%) OF DIFFERENT METHODS ON THE UCM-CAPTIONS DATASET

<table>
<thead>
<tr>
<th>Method</th>
<th>BLEU-1</th>
<th>BLEU-2</th>
<th>BLEU-3</th>
<th>BLEU-4</th>
<th>ROUGE-L</th>
<th>CIDEr-D</th>
</tr>
</thead>
<tbody>
<tr>
<td>VLAD+RNN [64]</td>
<td>63.11</td>
<td>51.93</td>
<td>46.06</td>
<td>42.09</td>
<td>58.78</td>
<td>200.66</td>
</tr>
<tr>
<td>VLAD+LSTM [67]</td>
<td>70.16</td>
<td>60.85</td>
<td>54.96</td>
<td>50.30</td>
<td>65.20</td>
<td>231.31</td>
</tr>
<tr>
<td>mRNN [67], [68]</td>
<td>60.10</td>
<td>50.70</td>
<td>32.80</td>
<td>20.80</td>
<td>-</td>
<td>214.00</td>
</tr>
<tr>
<td>mLSTM</td>
<td>63.50</td>
<td>53.20</td>
<td>37.50</td>
<td>21.30</td>
<td>-</td>
<td>222.50</td>
</tr>
<tr>
<td>mGRU</td>
<td>42.56</td>
<td>29.99</td>
<td>22.91</td>
<td>17.98</td>
<td>37.97</td>
<td>124.82</td>
</tr>
<tr>
<td>mGRU+embedword [78]</td>
<td>75.74</td>
<td>69.83</td>
<td>64.51</td>
<td>59.98</td>
<td>66.74</td>
<td>279.24</td>
</tr>
<tr>
<td>ConvCap [63]</td>
<td>70.34</td>
<td>56.47</td>
<td>46.24</td>
<td>38.57</td>
<td>59.62</td>
<td>190.15</td>
</tr>
</tbody>
</table>
### Table

<table>
<thead>
<tr>
<th>Method</th>
<th>L</th>
<th>M</th>
<th>S</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Soft-attention [67]</td>
<td>74.54</td>
<td>65.45</td>
<td>58.55</td>
<td>52.50</td>
</tr>
<tr>
<td>Hard-attention [67]</td>
<td>81.57</td>
<td>76.98</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CSMLF [69]</td>
<td>36.71</td>
<td>14.85</td>
<td>7.63</td>
<td>5.05</td>
</tr>
<tr>
<td>RTRMN (semantic) [71]</td>
<td>55.26</td>
<td>45.15</td>
<td>39.62</td>
<td>35.87</td>
</tr>
<tr>
<td>RTRMN (statistical) [71]</td>
<td>80.28</td>
<td>73.22</td>
<td>68.21</td>
<td>63.93</td>
</tr>
<tr>
<td>SAA</td>
<td>79.62</td>
<td>74.01</td>
<td>69.09</td>
<td>64.77</td>
</tr>
<tr>
<td>Baseline [72]</td>
<td>83.21</td>
<td>76.78</td>
<td>71.09</td>
<td>66.02</td>
</tr>
<tr>
<td>Structural representative network (ours)</td>
<td>84.22</td>
<td>78.25</td>
<td>73.25</td>
<td>68.62</td>
</tr>
</tbody>
</table>

* The “_” means that the scores are not reported in the reference papers.

### 4.4. Qualitative results

This section demonstrates the qualitative analysis of the presented structural representation network (SRN) through captions of each produced word for a given input image. It can be observed from the Figure that the presented approach is attending to the appropriate image section of each produced word. It can also be inferred that the produced caption is diverse, semantically descriptive, and precise. Particularly, the produced words determine that the presented approach is able to produce fine-grained words by using the holistic depiction of an image. Fig. 3 - Fig. 5. shows the captioning results of the baseline method and the proposed method on all the three datasets.
Fig. 3. Captioning results of the baseline method and the
proposed method on the UCM-Caption dataset [28].
5. CONCLUSION

The image captioning approach uses most of either attention-guided object/image-level features [59], [60], or visual entities [61], [62] for representing the gist of an image. However, these techniques fail to incorporate spatial details of small objects and multi-scale contextual details of images. To address this issue, a novel structural representation network (SRN) is proposed for image captioning. For caption generation tasks, the proposed SRN approach incorporates attention-guided dense visual features. Particularly, first, the spatial and semantic features are extracted from the backbone network. Further, a coarse-to-fine multi-stage caption decoder is used to produce fine-grained captions. Finally, the efficacy of the proposed approach is demonstrated on three remote sensing image captioning datasets, namely, UCM-Captions, Sydney-Captions, and RCISD. The stack of LSTM networks in the multi-stage caption decoder module incorporates intermediate supervision and handles the vanishing gradient problem.

6. REFERENCES


Neural image caption generation with visual attention,” in International conference on machine learning, pp. 2048-2057. PMLR, 2015.


**AUTHORS’ BACKGROUND**

<table>
<thead>
<tr>
<th>Your Name</th>
<th>Title*</th>
<th>Research Field</th>
<th>Personal website</th>
</tr>
</thead>
</table>

*This form helps us to understand your paper better, the form itself will not be published.*
*Title can be chosen from: master student, Phd candidate, assistant professor, lecture, senior lecture, associate professor, full professor