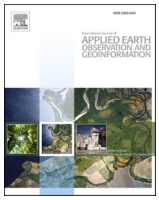
International Journal of Applied Earth Observations and Geoinformation 108 (2022) 102741

Contents lists available at ScienceDirect

International Journal of Applied Earth Observations and Geoinformation

journal homepage: www.elsevier.com/locate/jag



Transforming remote sensing images to textual descriptions

Usman Zia [a], M. Mohsin Riaz [b], Abdul Ghafoor [a,*]

[a] National University of Sciences and Technology (NUST), Pakistan
[b] COMSATS University Islamabad, Pakistan

ARTICLE INFO

Keywords:
Remote sensing image retrieval
Multi-modal domains
Image descriptions

ABSTRACT

Remote sensing data is growing enormously by virtue of the advancements in satellite and drone technology. Generating description of these remote sensing images have gained much attention in recent past owing to its applications in remote sensing image retrieval (RSIR) and image analysis. Remote sensing images capture huge diversity in different perspectives and levels. Due to overhead perspective and significantly larger scale of the scene, extraction of visual features from the remote sensing images to generate descriptions has become a challenging task. In this work, a model is proposed to generate novel captions by considering multi-scale features processed through adaptive attention based decoder using topic sensitive word embedding. The proposed model has been quantitatively evaluated using the benchmark remote sensing image captioning datasets and ablation study has been conducted to investigate the reason behind its effectiveness. Experimental evaluation depicts that the proposed model shows promising results compared to the existing state of the art remote sensing image description models.

1. Introduction

Remote sensing technologies have seen rapid development in recent past resulting in enhanced capabilities to capture geographic images in high resolution and fine details. With these advancements in remote sensing equipment and technologies, research on applications related to remote sensing images have gained much attention including image classification, object recognition and geographic scene segmentation (Ja and Raimond, 2015; Rahaman and Hassan, 2016; Ranzato et al., 2016). An interesting application of remote sensing image analysis is automatic generation of textual descriptions (Rennie et al., 2017). Automatic generation of description for remote sensing images have applications in both military and civilian sector. Image description models can generate textual information related to battlefield thereby enhancing information exchange between front line and command center. In civilian sector, textual descriptions of remote sensing images can boost image retrieval tasks (Hoxha et al., 2020). Image description task involves identification of visual features and their semantic relationship with natural language. Deep learning based automatic image description generation models learn the semantic relationship between the visual contents and corresponding textual attributes (Ling and Fidler, 2017). Using the learnt features, these models generate novel descriptions for unseen remote sensing images.

Remote sensing images capture huge diversity in different perspectives and levels (Wang et al., 2020a), in contrast to natural scene images. The natural scene images generally capture human perspective whereas remote sensing images capture overhead view of the terrain. Furthermore, remote sensing images are taken from a high altitude and capture a very large area as compared to natural images. Due to overhead perspective and significantly larger scale of the scene, objects are smaller and large in number. Therefore, extraction of visual features from the remote sensing images for generating descriptions becomes a challenging task.

Description generation models for remote sensing images proposed in recent past follow encoder-decode architecture (Wu et al., 2020a). These models break the description generation problem into two subtasks: encoder extracts visual features and learns their semantic relationship; while decoder utilizes these learnt features to generate a sequence of words representing human-like textual descriptions. Traditionally, Convolutional Neural Network (CNN) is utilized at the encoder stage to extract dominant visual features from the remote sensing images. Recurrent neural network with attention mechanism is employed at decoder stage to generate the textual descriptions. Backbone for extraction of visual feature used in recent remote sensing image description models mostly follow AlexNet (Krizhevsky et al., 2012), VGGNet(Simonyan and Zisserman, 2015), InceptionNets(Szegedy et al.,

* Corresponding author.
E-mail addresses: usman.phd@students.mcs.edu.pk (U. Zia), mohsin.riaz@comsats.edu.pk (M. Mohsin Riaz), abdulghafoor-mcs@nust.edu.pk (A. Ghafoor).

https://doi.org/10.1016/j.jag.2022.102741
Received 4 October 2021; Received in revised form 8 March 2022; Accepted 9 March 2022
Available online 28 March 2022

Fig. 1. An aerial image of airfield with distinct visual features at differnt scales.

2016) and ResNets(He et al., 2016). However, these models suffer from vanishing gradient and accuracy degradation problems (Heinrich et al., 2019). Furthermore, these models generally do not capture granular details peculiar to remote sensing images present at different scales. Fig. 1 shows a sample remote sensing image of an airfield. It can be observed that in order to capture all relevant details for generating a meaningful textual description, the visual attributes at different scale need to be considered.

Recently, attention mechanism is employed to identify relationship among different regions of images (Xu et al., 2015). Graph convolutional neural networks (Kipf and Welling, 2016) are utilized to relate regions in the image; however, in order to build the graph of visual features for text generation (Yao et al., 2018; Yang et al., 2019; Guo et al., 2019; Yao et al., 2019), axillary models are required in prior to guide attention. Transformer architecture(Vaswani et al., 2017) developed for natural language processing field relate word embedding of sentences without auxillary models and are trainable end to end. Image captioning models (Huang et al., 2019; Li et al., 2019; Herdade et al., 2019; He et al., 2020) proposed in recent past have adopted the transformer architectures to achieve state-of-the-art performance by implicitly relating informative regions in the image through dot-product attention.

Remote sensing image description models proposed in recent past have shown promising results, however there are still existing challenges such as: 1) Visual feature extraction is generally fixed in scale and semantic level. On the contrary, remote sensing images capture scenes with extremely brief (e.g. mountain range, forest) or highly intricate (e.g. highway interchange, airport) visual contents. It is difficult to extract meaningful attributes with scale and semantic level. 2) Existing models for remote sensing image description task do not cater for the polysemous nature of words while generating textual descriptions. This results in loss of essential semantic relationships among the sequence of words, making it difficult to generate more human like descriptions.

Proposed model addresses the challenges, including large variance in the visual aspects of objects caused by viewpoint variability, object obstruction, illumination, background clutter and shadow etc. The proposed model guides caption generation through multi-scale features. The adaptive multi-head attention decoder further refines the description generation by considering multi-scale features in addition to hidden states. In order to generate more human like and novel descriptions, the model incorporates topic sensitive word embedding(Zia et al., 2020) making it sensitive to polysemous nature of the words. Main contributions of the paper are:

- To cater for the large variation in remote sensing images due to scale and perspective, proposed model extracts multi-scale visual features for describing objects/scene at different scales.
- To generate human-like descriptions, adaptive multi-head attention based transformer decoder is proposed. The proposed adaptive decoder dynamically measures the contribution of multi-scale image features for next word prediction to generate fine grained descriptions.
- In language model, topic sensitive embedding is proposed to cater for polysemous nature of words.
- Evaluation and ablation study to validate performance of proposed model on state of the art remote sensing image description datasets.

2. Related work

2.1. Natural image description models

Image description models utilizing deep neural networks achieved state of the art performance due to the availability of large-scale annotated datasets (Deng et al., 2009; Chen et al., 2015a). These models generally follow encoder-decoder architecture for generating human-like image descriptions. Detailed visual attributes from the target image extracted by the encoder are used by the decoder based on language model for generating a sequence of words describing the visual contents as human-like descriptions. CNN based encoder and LSTM based decoder were first proposed by Vinyals et al. (Vinyals et al., 2015). Attention mechanism was added by Xu et al. (Xu et al., 2015) in the decoder stage. Their attention module assigns weights to each receptive filed of the encoded visual contents. The weighted matrix along with last generated word is then fed to the decoder to output the next word. Feature channel based attention model was proposed by Chen et al. (Chen et al., 2016) and Lu et al. (Lu et al., 2017a) proposed attention model to adaptively focus on visual contents more relevant to textual description.

Visual attention (Xu et al., 2015; Lu et al., 2017a; Chen et al., 2016) based models steer the caption generation process by focusing specific image features when computing word sequence. Detailed features, objects and attributes are fused with language model by You et al. (You et al., 2016). Adaptive attention model is proposed by (Gao et al., 2020)

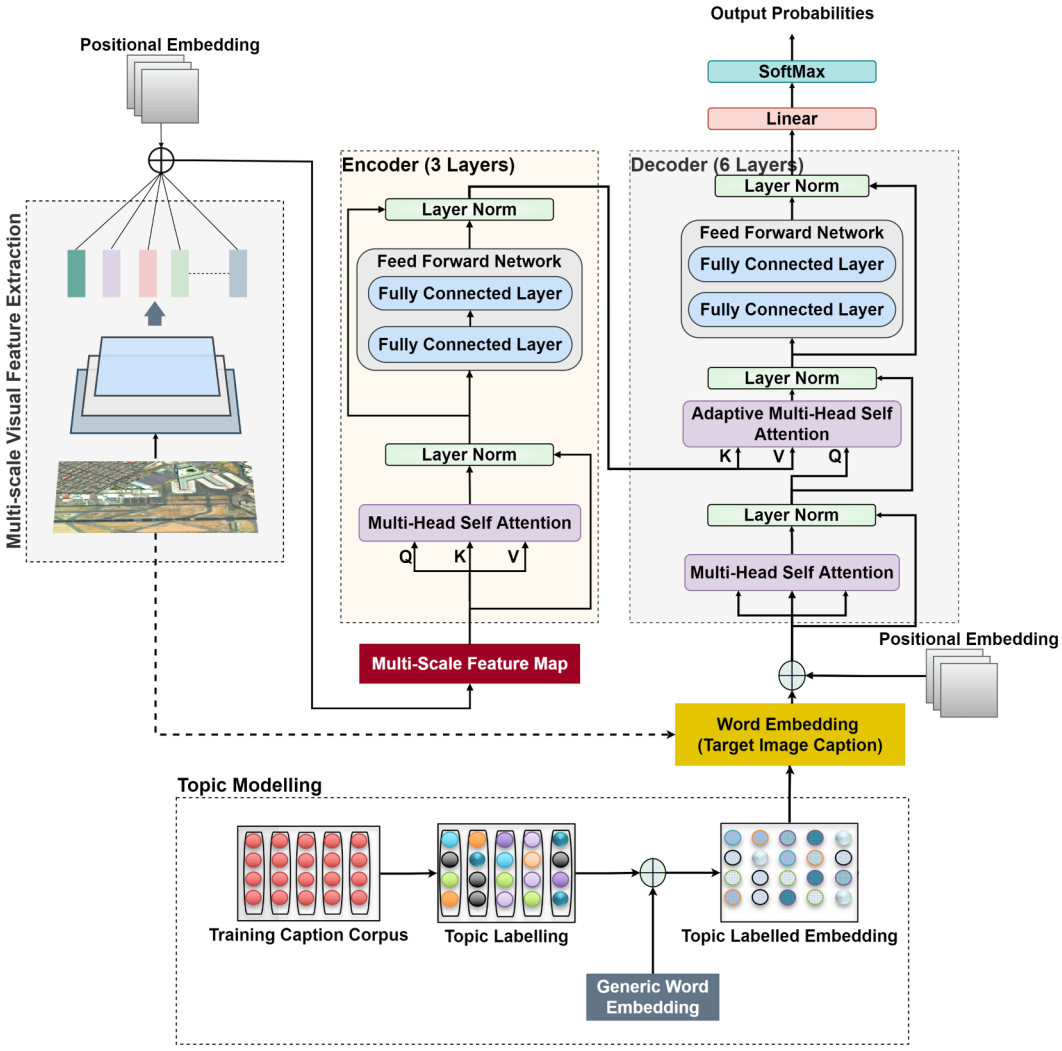


Fig. 2. Model Overview.

while model proposed by (Fang et al., 2015) boosts captioning process by incorporating object level details. Chen et al. (Chen et al., 2017) used R-LSTM to detect essential parts of image in prior for caption generation. Salient regions are used by Anderson et al. (Anderson et al., 2018) for attention to be calculated, however, the model neglects low-level visual attributes which are also useful for understanding the images. Gu et al. (Gu et al., 2018) consists of stacked coarse decoder and a series of fine decoders. The model randomly samples the intermediate outputs of the first decoder which are not well-defined and thus prone to accumulate errors and hard to train. Wang et al. (Wang et al., 2019b) applied a combination of attention mechanism using CNN and Faster R-CNN (Ren et al., 2015) to induce broad features and region proposals respectively to image captioning. However, relationship of each detected region with other regions is not considered during generation of description.

2.2. Remote sensing image description models

Generally, remote sensing image description generation methods are based on basic framework of natural image descriptions. Model proposed by Shi et al. (Shi and Zou, 2017) uses object detection based on fully convolutional network for extracting visual features for description generation without using LSTM. The model uses language generation template for description generation. Model proposed by Wang et al. (Wang et al., 2019a) uses retrieval based method to generate image descriptions using the distance of visual representations and sentence embedding.

Qu et al. (Qu et al., 2016) utilized encoder-decoder architecture by combining visual features extracted from remote sensing images and related sequence of words. CNN is used for extraction of visual embedding which are in turn fed to LSTM along with word embedding for description generation. Lu et al. (Lu et al., 2017b) extended the model by adding visual attention layer to the encoder-decoder model. Models following encoder-decoder architecture still struggle to generate optimum captions for remote sensing images owing to high similarly between the visual contents. To solve this problem, Zhang et al. (Zhang et al., 2019b) further refined the description generation model by adding an attribute attention mechanism based on low and high-level visual features. Wu et al. (Wu et al., 2020b) utilized mean feature of all CNN layers in addition to detailed features for effective caption generation. In contrast to traditional attention based models which use hidden state from previous time step, the model proposed by (Wu et al., 2020b) utilizes hidden state of current time step for computing attention maps. Wang et al. (Wang et al., 2020b) proposed remote sensing captioning method with scene attention mechanism. The model incorporates path and global features in addition to object level details for generation of attention mask to guide language model. However, model struggles in describing objects present at different scales. Model proposed by Zhang et al. (Zhang et al., 2019a) utilized traditional CNN to encode randomly cropped patches of remote sensing image and generate description using LSTM based language model. The attention mechanism used in these

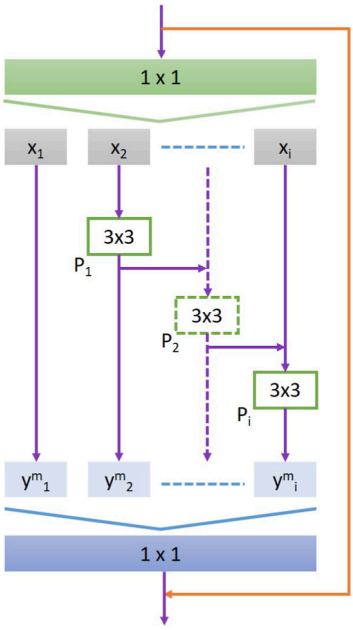


Fig. 3. Hierarchical Residual Block.

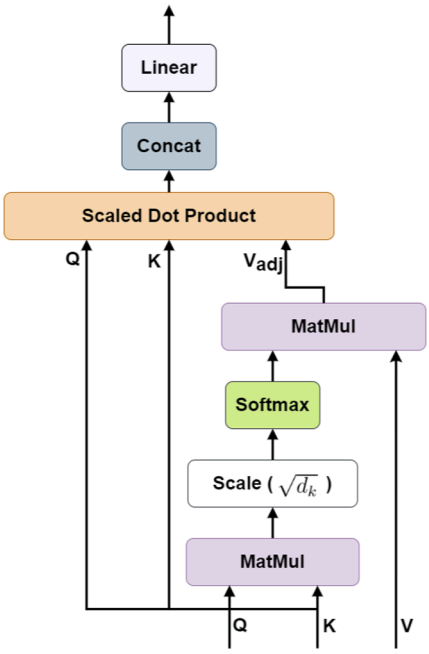


Fig. 5. Adaptive Attention Head.

models struggles to consider objects of interest present at different scales. Model proposed by (Lu et al., 2020) utilized sound-guided image features to generate novel captions. The model uses VGG-16 based visual feature extractor in the encoder stage augment by sound gated recurrent units (GRUs). The encoded image and sound features are fed to GRU based decoder to generate sequence of words. The model attends to different sounds associated with the image to obtain descriptions. (Shen et al., 2020) proposed remote sensing image captioning model utilizing CNN based visual features extractor and transformer decoder for word prediction. Transformer encoder and decoder comprises of many stacked identical layers in contrast to the recursive structure of LSTM. The decoder proposed in the model captures relationships in input sequence by incorporating residual connections around the stack structure. (Hoxha and Melgani, 2021) proposed image description model based on Support Vector Machine(SVM) classifier to generate the sentences. The model addressed expensive computational and training sample requirements of RNN based models by incorporating SVM at the decoder stage. (Sumbul et al., 2021) proposed summarization-driven captioning model. Caption generation is guided using summarized ground-truth captions focusing on relevant information. The model introduced summarization and fusion module as auxiliary components to provide flexible refinement to the encoder. The summarization module merges multiple single sentences generated by the encoder thus retaining the semantics of multiple sentences in a single caption. (Li

Table 1
Evaluation of proposed model using BLEU-N (N = 1,2,3,4), METEOR and CIDEr evaluation metrics on test images from RSICD dataset (Higher value represents better performance).

	RSICD					
Model	B1	B2	B3	B4	METEOR	CIDEr
(Wang et al., 2020b)	77.0	64.9	53.2	47.1	–	236.3
CCSMLF(Wang et al., 2019a)	57.6	38.6	28.3	22.2	21.3	53.0
Multimodal(Qu et al., 2016)	60.9	43.9	33.7	26.8	24.4	73.8
FC-ATT(Zhang et al., 2019b)	74.6	62.5	53.4	45.7	–	236.6
SM-ATT(Zhang et al., 2019b)	75.7	63.4	53.8	46.1	–	235.6
Scene-Attn(Wu et al., 2020c)	62.5	46.3	36.4	29.7	25.3	80.9
Sound-Active(Lu et al., 2020)	65.0	51.3	41.4	33.6	29.2	168.2
SCST(Shen et al., 2020)	77.0	**65.4**	56.3	48.8	**36.9**	**268.4**
PCE(Li et al., 2021)	74.4	61.8	52.9	46.1	34.2	235.2
SVM(Hoxha and Melgani, 2021)	61.1	42.7	31.5	24.1	23.0	68.2
Proposed Model	**79.8**	64.7	**56.9**	**48.9**	28.5	240.4

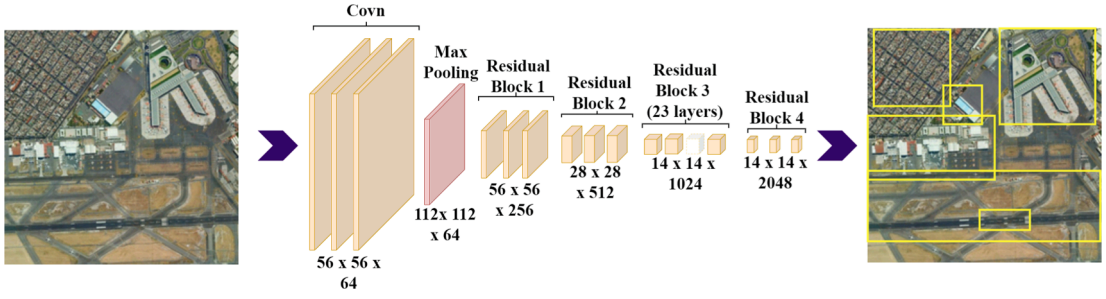


Fig. 4. Mulsti scale residual convolutional network.

Table 2
Evaluation of proposed model using BLEU-N (N = 1,2,3,4) evaluation metrics on test images from UCM dataset (Higher value represents better performance).

	UCM					
Model	B1	B2	B3	B4	METEOR	CIDEr
Multimodal(Qu et al., 2016)	78.7	71.0	64.9	59.4	40.4	292.7
(Zhang et al., 2019a)	59.4	53.2	48.1	42.9	–	–
CCSMLF(Wang et al., 2019a)	38.7	21.5	12.5	9.1	9.5	37.0
FC-ATT(Zhang et al., 2019b)	81.4	75.0	68.5	63.5	–	299.6
SM-ATT(Zhang et al., 2019b)	81.5	75.8	69.4	64.6	–	318.6
VAA (Zhang et al., 2019c)	81.9	75.1	69.2	63.8	43.8	339.4
Scene-Attn(Wu et al., 2020c)	82.2	76.5	71.7	67.4	44.0	322.8
(Wang et al., 2020b)	82.3	76.8	71.0	65.9	–	319.2
Sound-Active(Lu et al., 2020)	78.2	72.7	67.6	63.3	38.0	290.5
SCST(Shen et al., 2020)	83.8	**79.0**	**74.4**	**70.11**	44.6	**356.5**
SD-RSIC(Sumbul et al., 2021)	74.8	66.4	59.8	53.8	39.0	213.2
PCE(Li et al., 2021)	81.9	75.2	70.0	65.5	47.0	289.9
SVM(Hoxha and Melgani, 2021)	76.3	66.6	58.6	51.9	36.5	271.4
Proposed Model	**83.9**	76.9	71.5	67.5	**44.6**	323.1

et al., 2021) addressed model over-fitting problem by introducing truncation cross-entropy loss. Truncation cross entropy (TCE) based loss function proposed by the model utilized truncation threshold to reserve a margin for the non-target words and prevent over-optimization of the target words. Structured attention mechanism proposed by (Zhao et al., 2021) incorporates spatial relationships for generating location guided remote sensing image descriptions. The model uses regional information

Table 3
Evaluation of proposed model using BLEU-N (N = 1,2,3,4) evaluation metrics on test images from Sydney dataset (Higher value represents better performance).

	Sydney					
Model	B1	B2	B3	B4	METEOR	CIDEr
Multimodal(Qu et al., 2016)	71.5	60.8	52.9	46.2	34.8	195.6
(Zhang et al., 2019a)	61.5	54.0	47.3	40.0	–	–
CCSMLF(Wang et al., 2019a)	44.4	33.7	28.2	24.1	15.8	93.8
FC-ATT (Zhang et al., 2019b)	80.8	71.6	62.8	55.4	–	220.3
SM-ATT (Zhang et al., 2019b)	81.4	73.5	65.9	58.0	–	230.2
VAA (Zhang et al., 2019c)	74.3	66.4	60.2	54.9	39.3	240.7
(Wang et al., 2020b)	81.7	74.1	65.7	59.1	–	229.1
Scene-Attn (Wu et al., 2020c)	78.6	69.8	62.6	56.1	38.1	250.5
SCST(Shen et al., 2020)	80.1	70.9	63.1	57.2	41.8	253.9
Sound-Active(Lu et al., 2020)	71.5	63.2	54.6	46.6	31.2	180.2
SD-RSIC (Sumbul et al., 2021)	76.1	66.6	58.6	51.7	36.6	169.0
PCE (Li et al., 2021)	79.6	72.3	64.9	58.1	**42.4**	201.8
SVM(Hoxha and Melgani, 2021)	77.8	68.3	60.2	53.0	37.9	227.2
Proposed Model	**82.2**	**74.1**	**66.2**	**59.4**	39.7	**270.5**

Table 4
Comparison of Batch size, training parameters and training time.

Model	Batch Size	Parameters	Time
Soft-Attention (Lu et al., 2017b)	16	≈14 M	≈0.280s
m-Transformer (Shen et al., 2020)	16	≈28 M	≈0.185s
Proposed Model	16	≈29 M	≈0.186s

at pixel-level to enhance the efficacy of the proposed structured attention module. Structured unit obtained through the segmentation proposal generation module are utilized to exploit the spatial structure of semantic contents.

Availability of large captioning dataset for remote sensing images remained a challenge since the topic was rarely studied. Recently, Qu et al. (Qu et al., 2016) utilized UC Merced Land-Use(Yang and Shawn, 2010) and the Sydney dataset to propose UCM-Captions and Sydney-Captions dataset (Zhang et al., 2014). However, number of training images and scene categories in these two datasets are limited. RSICD dataset proposed by Lu et al. (Lu et al., 2017b) contains considerably greater number of images with higher intra-class diversity and lower inter-class similarity. The dataset is currently utilized as state of the art resource for tasks related to remote sensing image description.

3. Methodology

The proposed approach follows encoder-decoder architecture for description generation of remote sensing images. In this section, the overview of model is explained followed by detailed architecture of visual feature extractor in Section 3.2 and description generation in Section 3.3.

3.1. Model overview

Fig. 2 shows the network structure of the proposed model. Visual features are extracted from the multi-scale feature extractor. The topic labeled vocabulary is used to generate topic-sensitive word embedding to capture polysemous nature of words in training captions. Adaptive multi-head attention based decoder takes the visual features processed through the transformer encoder along with the topic sensitive embedding. The adaptive decoder considers visual features along with hidden states at each step for generating novel descriptions. Proposed visual feature extraction from multi-scale residual network is described in Section 3.2. Basic transformer architecture and proposed adaptive multi-head attention mechanism is 3.3, along with topic sensitive embedding of sequence of words generated form the training caption of the respective image.

3.2. Visual feature extraction

Remote sensing images capture granular details spread across large areas. Traditional visual feature encoders suffer from degradation problem (Heinrich et al., 2019) that results in reduced accuracy due to very deep architectures. (He et al., 2016) suggested to organize convolutional layers in residual blocks to avoid vanishing learning effects. However, feature extraction from remote sensing images requires a large range of receptive fields to describe objects at different scales. The proposed model extracts visual features at different scales to generate compact representation of the image contents. Visual extraction block of the proposed model generates visual features using efficient multi-scale processing method. Traditional ResNets (He et al., 2016) utilize an "identity shortcut connection" that skips one or more 3×3 convolutional layers represented as:

$$y = (x, W_i) + x \tag{1}$$

where input and output vector of the layers are represented by x and y. (x, W_i) depicts the residual mapping while operation $+x$ is performed by a shortcut connection and element-wise addition.

Instead of obtaining visual attributes using a group of 3×3 filters through the residual block, in (Gao et al., 2019), the visual features are extracted at much granular level. Multiple receptive fields are utilized at a more granular level to enhance the multi-scale representation ability. Proposed approach follows (Gao et al., 2019) to utilize c-channel based smaller groups of filters as hierarchical residual blocks instead of 3×3

GT:
The yacht is sailing in river with small houses at it.
Qu et al. (2016):
Many green trees are in two sides of a curved river.
Wang et al. (2020a):
In the middle of a curved bridge, of two green grasses of grassland areas in two rows of large river.
Proposed Model
A curved river is between the green fields.

GT:
A playground is built next to a white building.
Qu et al. (2016):
Many buildings and green trees are in a dense residential area.
Wang et al. (2020a):
There is a big playground in the school.
Proposed Model
A playground is next to residential area.

GT:
There are some white storage tanks near two lines of houses with trees.
Qu et al. (2016):
Some storage tanks are near a road.
Wang et al. (2020a):
In the open space in road, arranged are white storage tanks.
Proposed Model
White storage tanks are near some trees.

GT:
There is a viaduct with the shape of cloverleaf.
Qu et al. (2016):
Many green trees and some buildings are near a viaduct.
Wang et al. (2020a):
Several buildings are near a viaduct.
Proposed Model
Many trees are near a viaduct.

Fig. 6. Sample captions generated by the proposed model. Ground truth of random images from test split of RSICD dataset is compared with captions generated by proposed model. Novel but similar captions are generated by the proposed model by attending to details at differnt scales.

filters such that $n = \varsigma \times c$ where n in the number of channels in image (3 for coloured images) and ς is the scale dimension (4 in proposed approach). The number of scales to represent the output features are enhanced by connecting these smaller filter groups in a hierarchical residual-like style depicted in Fig. 3.

Feature map generated through 1×1 convolution is split into ς map subsets denoted by x_i , where $i \in 1, 2, \ldots, \varsigma$. Spatial size of each feature subset x_i is identical except number of channels is reduced to $1/\varsigma$. Each x_i is processed through a 3×3 convolution depicted by $P_i()$ to generate output z_i. Feature map subset x_i is added to the output of convolutional filter $P_{i-1}()$ before feeding into $P_i()$. The output can be expressed as:-

$$y_i^m = \begin{cases} x_i & i = 1 \\ P_i(x_i) & i = 2 \\ P_i(x_i + y_{i-1}^m) & 2 < i \leqslant s \end{cases} \quad (2)$$

Finally, the output y^m from multi-scale block is processed through 1×1 convolutional filters. The proposed model incorporates multi-scale hierarchical residual block within existing ResNet(He et al., 2016) backbone to extract visual features. Fine-tuning of (Gao et al., 2019) pre-trained on ImageNet is carried out on the benchmark remote sensing datasets including UCM(Qu et al., 2016) and Sydney(Zhang et al., 2014) datasets.

Let I denote the image of size $256 \times 256 \times 3$ processed through the pre-trained visual feature extraction backbone to obtain feature map (Fig. 4). Intermediate feature map of spatial size $14 \times 14 \times 2048$ is used to feed the subsequent language generation. The feature map is cut into 2048-dimensional 14×14 (196) vectors represented as $I_{FM} \in \mathbf{R}^{196 \times 2048}$ and fed to transformer block for generating textual descriptions.

3.3. Description generation

Natural Language Processing (NLP) models have been proposed in literature for generating n-dimensional embedding of words. Word2vec (Mikolov et al., 2013), GloVe(Pennington et al., 2014) and Elmo(Peters et al., 2018) have been utilized for generating semantic embedding. However, performance of these embedding techniques is not encouraging for polysemous words like *solution* (which can represent both *a mixture of liquids* and *a way to resolve a problem*) or *arms* (which can depict both a *part of body* or *a weapon*). These models struggle to capture contextual meaning and as a results embedding of polysemous words

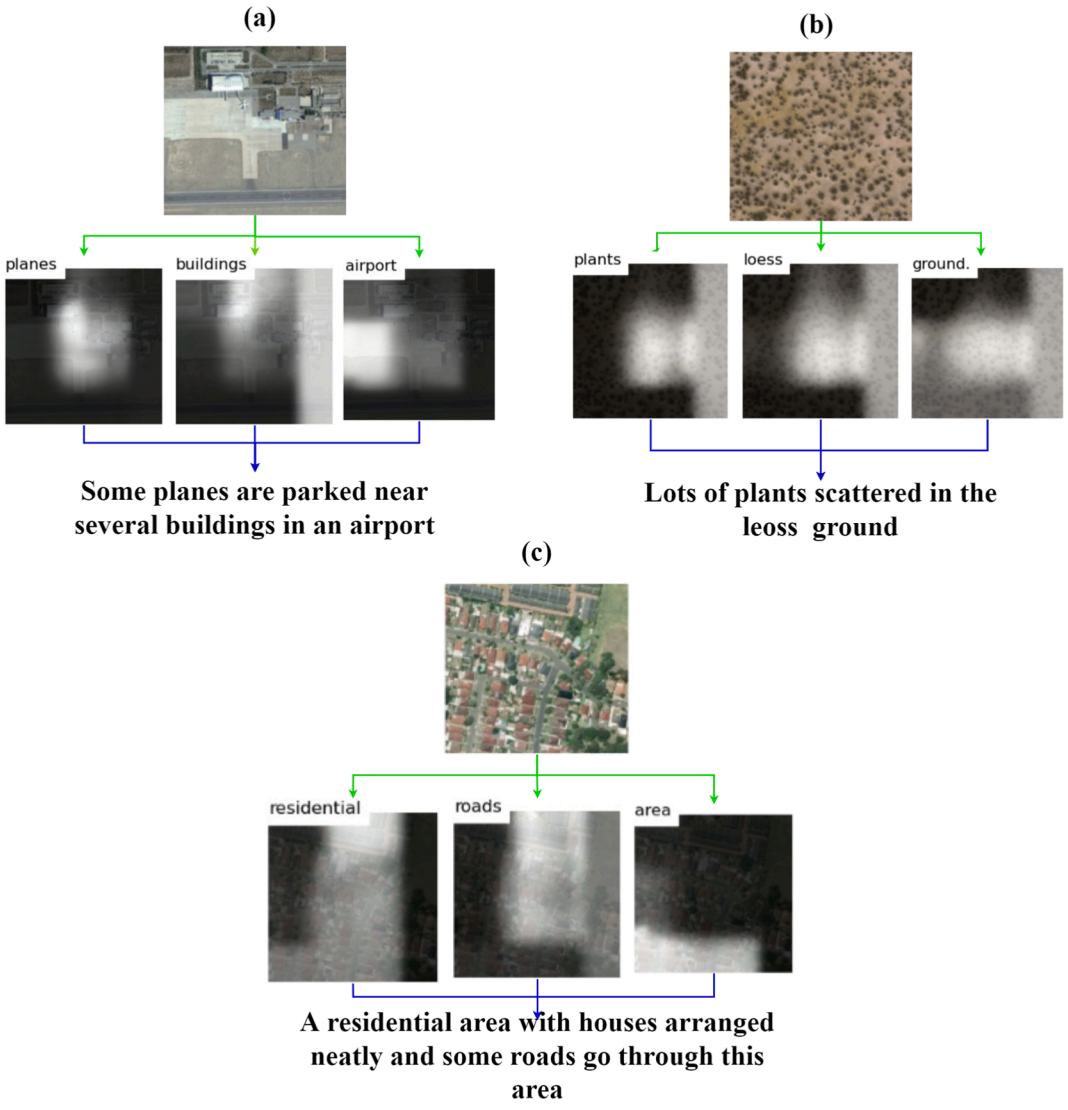


Fig. 7. Performance of proposed model to learn alignment between image patches and geenrated words.

like *apple* is represented by computing weighted average of *apple as a fruit*, *apple as an IT company* and *apple as a mobile device*. Topic models enable representation of topics as multinomial distribution over terms resultantly capturing polysemous nature of words. The proposed model utilizes topic modelling technique proposed in (Fadaee et al., 2017) for generating captions.

3.3.1. Textual embedding through topic modelling

Topic models assign a specific topic τ to each occurrence of word based on context. As a result, contextually similar occurrences of words are clustered together. Let $\mathscr{C}_I^j$ represent jth caption associated with Ith image in the training dataset where $\mathscr{C}_I$ represent the group of captions associated with the Ith image. The proposed model follows (Fadaee et al., 2017) to identify contextual topics of the words in vocabulary $\mathbb{V}$.

Captions associated with each image I are concatenated together and compiled as corpus $\mathbb{D}$. Corpus is processed through Hierarchical Dirichhilet Process (HDP) (Fadaee et al., 2017) to obtain topics τ_p associated with each word by computing the distribution over topics for words in vocabulary and for the caption group $\mathscr{C}_I$. As a result of processing, extended vocabulary $\acute{\mathbb{V}}$ now contains word-topic pairs as:

$$w \leftarrow - \; - \; \{v_l, \tau_p\} \tag{3}$$

The overall size of vocabulary is increased as a result of this operation since word v may form part of multiple word-topic pairs due to association with multiple topics τ based on the context.

Word embedding ϵ_j^i of (j) th word w_j belonging to caption C of Ith image is computed through vocabulary $\acute{\mathbb{V}}$ by maximizing the log-likelihood of the context words given word-topic pair:

$$\Gamma_{HTLE} = \frac{1}{N_{\acute{\mathbb{V}}}} \sum_{l=1}^{\mathbb{V}} \sum_{\substack{-c \leqslant d \leqslant n \\ j \neq 0}} \log p(v_{l \pm i} | v_l^{\tau}) \tag{4}$$

where $N_{\acute{\mathbb{V}}}$ represents the size of vocabulary and n_c depicts the number of context words considered for computing embedding.

Both visual features and topic based textual embedding are fed to transformer model for generating descriptions.

3.3.2. Language model

The goal of language model is to estimate probability of the next word in a sentence given the words and context seen previously. Proposed model takes the feature vector $I_{FM} \in \mathbf{R}^{196 \times 2048}$ from the visual encoder to generate a sequence of words as output. The feature vector is processed through scale-sensitive transformer composed of stacked encoder and adaptive decoder.

Fig. 8. Model tested on randomly selected images from RSICD test split by varying beam search k = 3, 5 and 7. Performance of model is found optimum with k = 5.

Traditional transformer-based image description models follow encoder-decoder framework. The encoder refines the visual representation by processing input from the visual feature extraction layer. The output word sequence is generated by the decoder using the encoded visual features. In proposed model, adaptive multi-head attention layer is introduced in the decoder. Visual feature vector is fed to the adaptive multi-head attention module to dynamically assign weights to multi-scale features and textual embedding during each step of the decoder.

Each encoder and decoder comprises of multiple layers (3 encoding and 6 decoding layers in proposed approach). Both encoder and decoder use self-attention to obtain the desired output (Fig. 2).

Attention Mechanism. Basic attention mechanism takes input i_n representing $n-th$ token from a set of N inputs. For each later attention layer, output of the preceding layer is fed to the next layer. Each attention block consists of a feed-forward neural network stacked on top of multi-head self-attention layer. Each of the h identical heads in multi-head attention layer computes the matrices query Q, key K and value V for each of the N tokens using:

$$Q = I_{FM}W_Q, K = I_{FM}W_K, V = I_{FM}W_V \tag{5}$$

where I_{FM} represents matrix of feature map while W_Q, W_K and $W_V \in \mathbf{R}^{h\times d_{feat}\times d_{attn}}$ are learnable projection matrices. h represents the number of attention heads, d_{feat} represents the dimension of image feature map and d_{attn} represents the feature dimension used to calculate attention weights. In the proposed approach, h is set to 8, d_{feat} as 2048 and d_{attn} as 64. Attention weight Ψ_K^Q is computed as dot product of Query Q and Key K matrix scaled with d_{attn} :

$$\Psi_K^Q = \frac{QK^T}{\sqrt{d_{attn}}} \tag{6}$$

where each element ψ^{mn} of $N \times N$ attention weight matrix Ψ_K^Q depicts the attention weight between n-th and m-th token. Dimension $d_attn = 64$ of the key, query, and value vectors is used as scaling factor for the computation.

The output of each attention head, representing self-attention, is obtained by multiplying attention weights Ψ_K^Q with the values V

$$Head(I_{FM}^j) = self-attention(Q,K,V) = softmax(\Psi_K^Q)V \tag{7}$$

where j depicts each attention head.

All multi-head attention results are concatenated along the h dimension and projected linearly to obtain the final attention output:

$$Attn(Q,K,V) = concat(Head(I_{FM}^1),\ldots,Head(I_{FM}^j)).W_O \tag{8}$$

where $W_O \in R^{d_{feat}\times d_{feat}}$.

Each output from the attention layer is finally used as input for the

Table 5
Ablation study of proposed approach using BLEU-N (N = 1,2,3,4), METEOR and CIDEr evaluation metrics on test images from RSICD (Higher value represents better performance).

Model	RSICD Dataset					
	B1	B2	B3	B4	METEOR	CIDEr
VGG16 + Trans	78.2	63.1	48.1	35.9	27.9	114.8
VGG16 + Trans + Topic	78.9	63.5	48.5	36.2	28.4	130.2
Multi Scale + Trans	79.5	63.9	49.3	36.7	28.5	223.1
Proposed Model	**79.8**	**64.7**	**56.9**	**48.9**	**28.5**	**240.4**

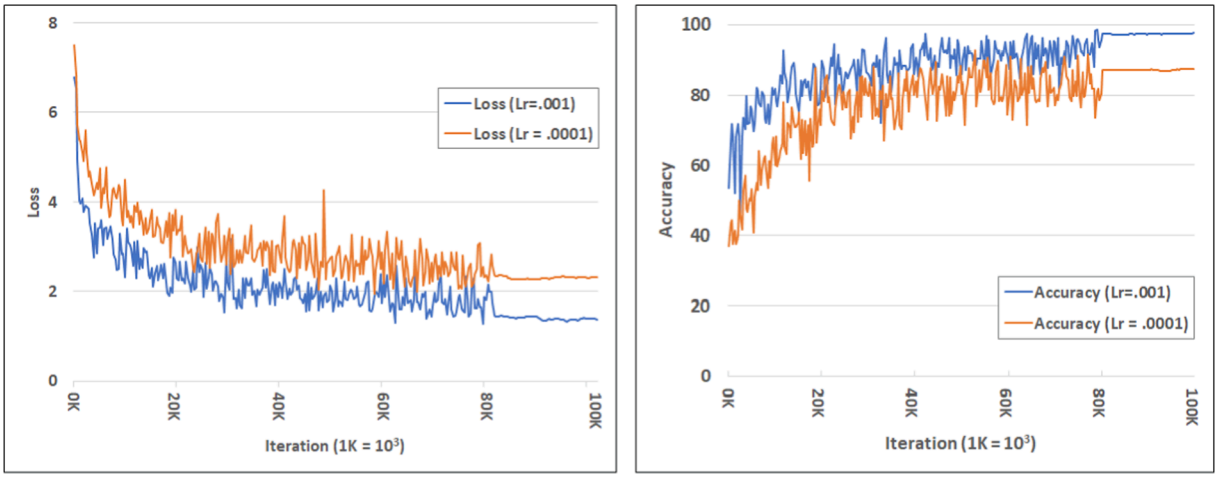


Fig. 9. Comparison of training loss and accuracy of the proposed model for different learning rates.

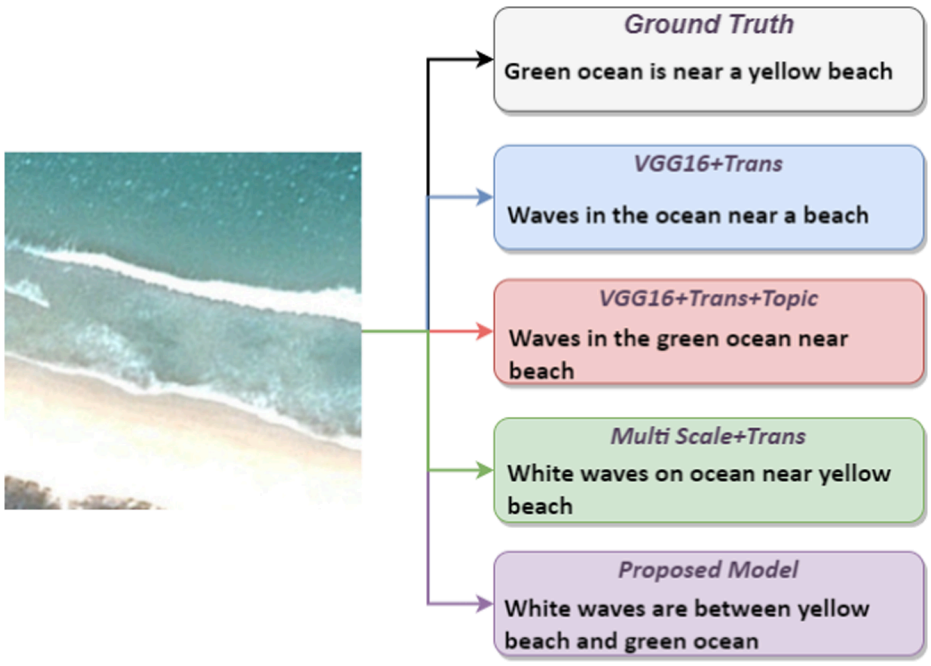


Fig. 10. Performance of Proposed Model in comparison to ablated version.

feed-forward network with a *Relu* activation function followed by layer normalization:

$$FeedForward(Attn) = Relu(Attn \cdot W_1 + b_{b1}) \cdot W_2 + b_2 \tag{9}$$

$$FeedForwardOut = LayerNorm(FeedForward(Attn) + Attn) \tag{10}$$

where W_1, W_2, b_{b1} and b_2 are learnable parameters.

Encoder. The encoder contains three stacked self-attention layers as explained in Section 3.3.2 and elaborated in Fig. 2. Transformer encoder, by design, cannot track the input order. Therefore, prior to feeding the visual features to the encoder, positional embedding is required. The proposed approach follows (Vaswani et al., 2017) to generate positional embedding P. The encoder is stacked with 3 self attention layers and the output of third self-attention layer is fed to the decoder during inter-attention.

Decoder. Topic based textual embedding of the target caption along with positional encoding and encoder output is fed to the decoder. Proposed model comprises of 6 stacked self-attention layers followed by softmax to obtain the final word predictions. Regions in different scales have different impact on the caption generation. In order to guide caption generation based on scale information computed by the multi-scale feature extractor, visual features are fed to the decoder during each time step. Prior to feeding encoded features to the decoder, weights are assigned dynamically to the multi-scale features (Fig. 5). The adjusted feature vector V_{adj} is computed by scaling V by the correlation between Q and K matrices as follows:-

$$V_{adj} = softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{11}$$

where $\sqrt{d_k}$ is used to ensure valid range of the output.

The adjusted features are in turn fed to the decoder for computing adaptive attention using Eq. 7:-

$$Head(I^j_{FM}) = self - attention(Q, K, V_{adj}) = softmax(\Psi^Q_K)V_{adj} \tag{12}$$

In order to achieve parallelism in the decoder, instead of feeding one generated word at a time, actual target sequence embedding is fed during the training phase. However, to restrict self-attention to only attend to words generated so far, attention-masking is applied.

The next word is predicted using the output feature of the last decoder layer whose output dimension equals to the vocabulary size. The proposed model minimizes the cross entropy loss function as under:

$$L(\theta) = \sum_{T=0}^{T} log(p_\theta(y^*_t | y^*_{0:t-1}, I_{FM})) \tag{13}$$

where θ denotes the model parameters and $y^*_0, y^*_1 \ldots y^*_t$ represent the target ground-truth sequence of words.

4. Results

Evaluation of the proposed approach is carried out on existing state of the art remote sensing datasets including RSICD(Lu et al., 2017b), UCM-captions(Qu et al., 2016) and Sydney-captions(Zhang et al., 2014).

4.1. Datasets

Quantitative evaluation of the model is carried out on state-of-the-art remote sensing image caption dataset. 80% : 10% : 10% split is used for training, validation and test sets respectively.

4.1.1. UCM-Captions(Qu et al., 2016)

UCM-Captions was originally compiled for scene classification task and was later extended with manual descriptions to be used for caption generation. Dataset contains 256 × 256 sized 2100 images classified into 21 categories with 100 images per category and 5 descriptions with each image.

4.1.2. Sydney-caption(Zhang et al., 2014)

Sydney-caption is an extended scene classification dataset for remote sensing image description tasks. The dataset is small-scale and comprises of only 613 images of size 500 × 500 segregated into 7 categories with 5 descriptions each.

4.1.3. RSICD(Lu et al., 2017b)

RSICD is largest remote sensing image description dataset containing 10,921 images. However, dataset contains an average 2–3 captions per image with only 24,333 ground truth descriptions. Existing sentences

are replicated to satisfy the input requirements of 5 captions per image.

4.2. Data augmentation

Dataset currently available for remote sensing image captioning have very limited number of training and validation images. RSICD (Lu et al., 2017b) although being the largest dataset for image captioning has only 10 K images which might result in over-fitting. In order to overcome this problem, data augmentation techniques have been employed to effectively enhance the dataset and alleviate over-fitting problem. Random mirroring in horizontal or vertical direction and/ or 45^o rotation is applied to each training image to augment the dataset. This results in 4 additional images with different rotation and flip. As a result of augmentation process, training dataset size is increased four folds for RSICD (Lu et al., 2017b), UCM(Qu et al., 2016) and Sydney(Zhang et al., 2014) dataset.

4.3. Implementation details

Each input image is resized to 256×256 before processing through visual feature extraction. Fine image features for the augmented training images are extracted through fine-tuned multi-scale residual network. Topic labelled words from all captions are sorted as per the frequency of their appearance to form vocabulary of the training dataset. Transformer encoder and decoder network is implemented in PyTorch framework. Adam optimizer(Kingma and Ba, 2014) is used for optimization and by reducing the cross-entropy loss between predicted and ground truth with a learning rate of $l = 0.001$. Dropout with probability $p = 0.1$ is used as regularization method. The model is trained for 60 epoch on a single NVIDIA RTX 3080 Ti GPU for 37 h with batch size of 16. Captions generated by the model are restricted to maximum length of 15 words. Early-stopping technique is employed to break the training iteration by monitoring validation loss convergence with patience of 3 epochs. Beam search is adopted to select best k generated words at a given time step for generation of next word. $k = 5$ is used for evaluation of proposed model.

4.4. Evaluation metrics

Performance of proposed model is evaluated for description generation using state of the art evaluation metrics with coco-caption(Chen et al., 2015b) code as follows.

- BLEU-N (Bilingual Evaluation Understudy with N = 1,2,3,4) gauges the precision of an n-gram between the computed and ground truth captions. The scores are computed using:

$$\log(B_N) = \min(1 - \frac{GT}{CP}, 0) + \sum_{n=1}^{N} \omega_n \log(p_n) \tag{14}$$

  Where GT and CP are the lengths of ground truth and the computed caption respectively, ω_n is the weight while p_n depicts the precision.
- METEOR (Metric for Evaluation of Translation with Explicit Ordering) is used as an evaluation metric in machine translation. The score generated by METEOR uses recall between the reference sentences and the computed captions in addition to precision.
- CIDEr (Consensus-based Image Description Evaluation) computes the similarity of computed captions to ground truth sentences by converting each vocabulary word to it's root form and taking into account the grammaticality and correctness.

4.5. Performance comparison

Proposed model shows competitive performance in comparison with existing state-of-the-art research for test images from RSICD, UCM and Sydney datasets. Table 1–3 show results on RSICD, UCM and Syndey datasets respectively.

In this work, experiments of the proposed model have been conducted on Ubuntu 16.0 LTS with Intel (R) Core(TM) i7-8700 at 3.20 GHz and NVIDIA RTX 3080 Ti GPU. Table 4 shows the comparison of batch size, number of training parameters and training time of the proposed approach with traditional LSTM based attention model and transformer based captioning model. Proposed model takes about 0.186s per iteration with batch size set to 16. Training time per parameter of the proposed model is comparable to existing state of the art image captioning models based on transformer architecture.

4.6. Qualitative evaluation

The performance of model is evaluated in generating human-like captions by considering multi-scale image features generated by residual network and processed through transformer based encoder-decoder. Fig. 6 compares the captions generated by the proposed model with ground truth for random images selected form test split of the RSICD dataset.

The proposed model learns alignment between image patches and words that strongly agree with human intuition. Fig. 7 shows the visualization of words-to-patch cross attention weights during the caption generation process. The model is able to correctly attend to relevant patches in the image for outputting next word in the sequence.

4.7. Parameter effect

4.7.1. Beam search

Max probability and beam search are two most common word sampling algorithms used in NLP tasks. The max probability sampling adopts a greedy strategy with less time and space complexity. However, the approach struggles to determine the global optimal solution. Beam search follows heuristic search algorithm by retaining k number of words (denoted as beam size) at each time step with the highest probability. Although, beam search is computationally expensive, however, the algorithm is able to determine optimal solution. The model is evaluated by varying the beam size k to observe the effect on quality of generated captions. 1000 randomly selected images from test split of *RSICD* dataset are processed through the model by setting k to 3, 5 and 7. It is observed that $k = 5$ entails optimum results as evident from captions mentioned in Fig. 8.

4.7.2. Learning rate

The impact of learning rate is analyzed on the training loss and accuracy for the proposed model. Fig. 9 shows that the model converges faster at learning rate $lr = 0.001$ compared to $lr = 0.0001$. Setting the value too low delays the learning process due to very minute update during each iteration. Similar behavior is observed in model accuracy. The number of iterations required to reach the same accuracy of the model are significantly less for learning rate $lr = 0.001$ compared to $lr = 0.0001$.

4.8. Ablation study

The performance of the proposed model is evaluated against different ablated versions on RSICD dataset. Beam search is fixed to 5 and batch size to 16 for all ablated versions. Model (VGG16 + Trans) using traditional CNN (VGG16) for visual encoder and Transformer as language generation model is used as baseline depicted in the first row of Table 5. Second row shows the results of model (VGG16 + Trans + Topic) based on Baseline augmented with topic modelling. VGG16 visual features extractor is replaced with multi-scale encoder in ablated version (Multi Scale + Trans) depicted in third row. Results of proposed model (Multi Scale + Trans + topic) depicted in fourth row incorporate multi-scale feature encoder with adaptive transformer decoder and topic modelling. Results show that by incorporating multi-scale features

processed through adaptive decoder in caption generation, significant improvement is achieved. Furthermore, the results reveal that using topic modelling along with multi-scale features indeed improve the performance.

Fig. 10 shows the visual comparison between the ablated versions and the proposed model. The description generated by the proposed model is able to represent finer details at different scales of the image compared to the ablated versions.

5. Conclusion

Proposed model addresses the challenges in remote sensing image description due to large variance in the visual aspects of objects. The proposed model extracts detailed information from remote sensing images through multi-scale visual feature encoder. Adaptive attention decoder dynamically assigns weights to the multi-scale features and textual queues to strengthen the language model to generate novel topic sensitive descriptions. The model shows promising performance on three benchmark remote sensing image description datasets. Ablation study and experimental evaluation depicts the effectiveness of the proposed model compared to the existing state of the art models. Future research directions include application of proposed model to generate description for different domains such as medical images.

CRediT authorship contribution statement

Usman Zia: Conceptualization, Methodology, Software, Writing-Original draft preparation. **Mohsin Riaz**: Data curation, Investigation, Visualization, Validation. **Abdul Ghafoor** : Writing- Reviewing and Editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., Zhang, L., June 2018. Bottom-up and top-down attention for image captioning and visual question answering. In: Proceedings of CVPR, pp. 6077–6086.

Chen, L., Zhang, H., Xiao, J., Nie, L., Shao, J., Liu, W., Chua, T.-S., 2016. Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning.

Chen, M., Ding, G., Zhao, S., Chen, H., Liu, Q., Han, J., 2017. Reference based lstm for image captioning. AAAI Conference 31, 3981–3987.

Chen, X., Fang, H., Lin, T.Y., Vedantam, R., Gupta, S., Dollár, P., Zitnick, C.L., 2015a. Microsoft coco captions: Data collection and evaluation server, arXiv preprint.

Chen, X., Fang, H., Lin, T.Y., Vedantam, R., Gupta, S., Dollár, P., Zitnick, C.L., 2015b. Microsoft coco captions: Data collection and evaluation server, arXiv preprint.

Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L., 2009. Imagenet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. pp. 248–255.

Fadaee, M., Bisazza, A., Monz, C., 2017. Learning topic-sensitive word representations. Annual Meeting of the Association for Computational Linguistics, pages. Asso. Comput. Linguist. 55, 441–447.

Fang, H., Gupta, S., Iandola, F., Srivastava, R., Deng, L., Dollar, P., Gao, J., He, X., Mitchell, M., Platt, J., Zitnick, C.L., Zweig, G., 2015. From captions to visual concepts and back. In: IEEE Conference on Computer Vision and Pattern Recognition.

Gao, L., Li, X., Song, J., Shen, H.T., 2020. Hierarchical lstms with adaptive attention for visual captioning. IEEE Trans. Pattern Anal. Mach. Intell. 42 (5), 1.

Gao, S., Cheng, M.M., Zhao, K., Zhang, X.Y., Yang, M.H., Torr, P.H., 2019. Res2net: A new multi-scale backbone architecture. IEEE Trans. Pattern Anal. Machine Intell.

Gu, J., Cai, J., Wang, G., Chen, T., 2018. Stack-Captioning: Coarse-to-Fine Learning for Image Captioning. In: AAAI.

Guo, L., Liu, J., Tang, J., Li, J., Luo, W., Lu, H., 2019. Aligning linguistic words and visual semantic units for image captioning. In: Proceedings of the 27th ACM International Conference on Multimedia. pp. 765–773.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR. pp. 770–778.

He, S., Liao, W., Tavakoli, H.R., Yang, M., Rosenhahn, B., Pugeault, N., 2020. Image captioning through image transformer. In: Asian Conference on Computer Vision.

Heinrich, K., Janiesch, C., Möller, B., Zschech, P., 12 2019. Is bigger always better? lessons learnt from the evolution of deep learning architectures for image classification. In: Pre-ICIS SIGDSA Symposium.

Herdade, S., Kappeler, A., Boakye, K., Soares, J., 2019. Image captioning: Transforming objects into words. Adv. Neural Informat. Process. Syst. 11135–11145.

Hoxha, G., Melgani, F., 2021. A novel svm-based decoder for remote sensing image captioning. IEEE Trans. Geosci. Remote Sens.

Hoxha, G., Melgani, F., Demir, B., 2020. Toward remote sensing image retrieval under a deep image captioning perspective, 4462–4475.

Huang, L., Wang, W., Chen, J., Wei, X.Y., 2019. Attention on attention for image captioning. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 4634–4643.

Ja, R., Raimond, K., 2015. A review on availability of remote sensing data. IEEE Technological innovation in ICT for agriculture and rural development (TIAR.

Kingma, D.P., Ba, J., 2014. Adam: A method for stochastic optimization, arXiv preprint.

Kipf, T.N., Welling, M., 2016. Semi-supervised classification with graph convolutional networks, arXiv preprint.

Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks. Adv. Neural Informat. Process. Syst. 1097–1105.

Li, G., Zhu, L., Liu, P., Yang, Y., 2019. Entangled transformer for image captioning. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 8928–8937.

Li, X., Zhang, X., Huang, W., Wang, Q., 2021. Truncation cross entropy loss for remote sensing image captioning. IEEE Trans. Geosci. Remote Sens. 5246–5257.

Ling, H., Fidler, S., 2017. Teaching machines to describe images via natural language feedback. In NIPS.

Lu, J., Xiong, C., Parikh, D., Socher, R., July 2017a. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In: Processing of IEEE Conference on Computer Vision and Pattern Recognition (CVPR. pp. 375–383.

Lu, X., Wang, B., Zhen, X., 2020. Sound active attention framework for remote sensing image captioning. IEEE Trans. Geosci. Remote Sens. 1985–2000.

Lu, X., Wang, B., Zheng, X., Li, X., 2017b. Exploring models and data for remote sensing image caption generation. IEEE Trans. Geosci. Remote Sens. 2183–2195.

Mikolov, T., Chen, K., Corrado, G., Dean, J., 2013. Efficient estimation of word representations in vector space. ICLR Workshop.

Pennington, J., Socher, R., Manning, C., 2014. Glove: Global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). pp. 1532–1543.

Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., Zettlemoyer, L., 2018. Deep contextualized word representations. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 2227–2237.

Qu, B., Li, X., Tao, D., Lu, X., 2016. Deep semantic understanding of high resolution remote sensing image. In: 2016 International Conference on Computer. Information and Telecommunication Systems (CITS).

Rahaman, K.R., Hassan, Q.K., 2016. Application of remote sensing to quantify local warming trends: A review. In: 5th International conference on informatics. electronics and vision, pp. 256–261.

Ranzato, M., Chopra, S., Auli, M., Zaremba, W., 2016. Sequence level training with recurrent neural networks. In: 4th International conference on learning representations. ICLR.

Ren, S., He, K., Girshick, R., Sun, J., 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems.

Rennie, S.J., Marcheret, E., Mroueh, Y., Ross, J., Goel, V., 2017. Self-critical sequence training for image captioning. In: 30th IEEE conference on computer vision and pattern recognition (CVPR 2017. pp. 1179–1195.

Shen, X., Liu, B., Zhou, Y., Zhao, J., 2020. Remote sensing image caption generation via transformer and reinforcement learning. Multimedia Tools Appl. 26661–26682.

Shi, Z., Zou, Z., 2017. Can a machine generate humanlike language descriptions for a remote sensing image? IEEE Trans. Geosci. Remote Sens. 3623–3634.

Simonyan, K., Zisserman, A., 2015. Very deep convolutional networks for large-scale image recognition. ICLR.

Sumbul, G., Nayak, S., Demir, B., 2021. Sd-rsic: Summarization-driven deep remote sensing image captioning. IEEE Trans. Geosci. Remote Sens. 6922–6934.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z., 2016. Rethinking the inception architecture for computer vision. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR. pp. 2818–2826.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I., 2017. Attention is all you need. Adv. Neural Informat. Process. Syst. 5998–6008.

Vinyals, O., Toshev, A., Bengio, S., Erhan, D., 2015. Show and tell: A neural image caption generator. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR),, MA. pp. 3156–3164.

Wang, B., Lu, X., Zheng, X., Li, X., 2019a. Semantic descriptions of high-resolution remote sensing images. IEEE Geosci. Remote Sens. Lett.

Wang, C., Jiang, Z., Yuan, Y., 2020a. Instance-aware remote sensing image captioning with cross-hierarchy attention. In: IEEE International Geoscience and Remote Sensing Symposium, pp. 980–983.

Wang, C., Jiang, Z., Yuan, Y., 2020b. Instance-aware remote sensing image captioning with cross-hierarchy attention. IGARSS (2020 IEEE International Geoscience and Remote Sensing Symposium) 2020, pp. 980–983.

Wang, E.K., Zhang, X., Wang, F., Wu, T.Y., Chen, C.M., 2019b. Multilayer dense attention model for image caption. IEEE Access.

Wu, S., Zhang, X., Wang, X., Li, C., Jiao, L., 2020a. Scene attention mechanism for remote sensing image caption generation. In: 2020 International Joint Conference on Neural Networks (IJCNN. pp. 1–7.

Wu, S., Zhang, X., Wang, X., Li, C., Jiao, L., 2020b. Scene attention mechanism for remote sensing image caption generation. In: 2020 International Joint Conference on Neural Networks (IJCNN). pp. 1–7.

Wu, S., Zhang, X., Wang, X., Li, C., Jiao, L., 2020c. Scene attention mechanism for remote sensing image caption generation. In: 2020 International Joint Conference on Neural Networks (IJCNN). pp. 1–7.

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhutdinov, R., Zemel, R., Bengio, Y., 2015. Show, attend and tell: Neural image caption generation with visual attention. In: International Conference on Machine Learning.

Yang, X., Tang, K., Zhang, H., 2019. H and j. In: Auto-encoding scene graphs for image captioning. In: Cai, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 10685–10694.

Yang, Y., Shawn, N., 2010. Bag-of-visual-words and spatial extensions for land-use classification. In: 18th SIGSPATIAL international conference on advances in geographic information systems.

Yao, T., Pan, Y., Li, Y., Mei, T., 2018. Exploring visual relationship for image captioning. In: Proceedings of the European Conference on Computer Vision (ECCV. pp. 684–699.

Yao, T., Pan, Y., Li, Y., Mei, T., 2019. Hierarchy parsing for image captioning. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2621–2629.

You, Q., Jin, H., Wang, Z., Fang, C., Luo, J., 2016. Image captioning with semantic attention. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR. pp. 4651–4659.

Zhang, F., Du, B., Zhang, L., 2014. Saliency-guided unsupervised feature learning for scene classification. IEEE Trans. Geosci. Remote Sens. 2175–2184.

hang, X., Wang, Q., Chen, S., Li, X., 2019a. Multi-scale cropping mechanism for remote sensing image captioning. In: IEEE International Geoscience and Remote Sensing Symposium, pp. 10039–10042.

Zhang, X., Wang, X., Tang, X., Zhou, H., Li, C., 2019b. Description generation for remote sensing images using attribute attention mechanism. Remote Sensing.

Zhang, Z., Zhang, W., Diao, W., Yan, M., Gao, X., Sun, X., 2019c. Vaa: Visual aligning attention model for remote sensing image captioning. IEEE Access 137355–137364.

Zhao, R., Shi, Z., Zou, Z., 2021. High-resolution remote sensing image captioning based on structured attention. IEEE Trans. Geosci. Remote Sens. 60, 1–14.

Zia, U., Riaz, M.M., Ghafoor, A., Ali, S.S., 2020. Topic sensitive image descriptions. Neural Comput. Appl. 1–9.