Indonesian Journal of Electrical Engineering and Informatics (IJEEI)
Vol. 11, No. 1, March 2023, pp. 267~285
ISSN: 2089-3272, DOI: 10.52549/ijeei.v11i1.4342

Automatic Caption Generation for Aerial Images: A Survey

Parag J. Mondhe[1], Manisha P. Satone[2], Gajanan K. Kharate[3]
[1,2,3]Department of Electronics and Telecommunication Engineering, Matoshri College of Engineering and Research Centre, Nashik, India affiliated to Savitribai Phule Pune University, Pune, India.

Article Info

Article history:

Received Nov 22, 2022
Revised Mar 4, 2023
Accepted Mar 19, 2023

Keyword:

Aerial Images
Caption Generation
Description Generation
Remote Sensing Images
Satellite Images

ABSTRACT

Aerial images have attracted attention from researcher community since long time. Generating a caption for an aerial image describing its content in comprehensive way is less studied but important task as it has applications in agriculture, defence, disaster management and many more areas. Though different approaches were followed for natural image caption generation, generating a caption for aerial image remains a challenging task due to its special nature. Use of emerging techniques from Artificial Intelligence (AI) and Natural Language Processing (NLP) domains have resulted in generation of accepted quality captions for aerial images. However, lot needs to be done to fully utilize potential of aerial image caption generation task. This paper presents detail survey of the various approaches followed by researchers for aerial image caption generation task. The datasets available for experimentation, criteria used for performance evaluation and future directions are also discussed.

Corresponding Author:

Parag J. Mondhe,
Department of Electronics and Telecommunication Engineering,
Matoshri College of Engineering and Research Centre, Nashik, India affiliated to Savitribai Phule Pune University, Pune, India.
Email: mondheparag@gmail.com

1. INTRODUCTION

An image can be best understood by its caption. Images can be taken from front, side or top. An aerial image is taken from top using either Unmanned Aerial Vehicle (UAV) which is also referred as Unmanned Aerial System (UAS), or using either drone, helicopter, or satellite. The caption for an aerial image describes it in such a way that main objects and their inter-relationship can be captured in a sentence. Aerial images are useful in numerous applications, however most of the work was devoted to object detection [1-7] or scene classification [8-14] in the past. Generating a caption is different than only detecting objects from an image. To generate caption, knowledge of Natural Language Processing (NLP) is required along with image processing and pattern recognition. In caption generation, attributes of the objects and their relationships needs to be understood and it should be comprehensively described using natural language. Moreover, the generated caption must be meaningful and consistent with the human language. Aim of this paper is to extensively survey the literature associated with automatic caption generation algorithms for aerial images.

This paper is organized as follows: section 2 explores the applications of generating caption for aerial images, section 3 discussed the challenges in dealing with aerial images. Section 4 provides survey of the approaches followed by researchers for aerial image caption generation. Section 5 describes the datasets available and used by different researchers while section 6 explains the evaluation metrics used to assess quality of generated captions. Section 7 provides future directions for the aerial image caption generation task.

2. APPLICATIONS OF GENERATING CAPTION FOR AERIAL IMAGES

Though generating caption for aerial images is a challenging task, it has numerous applications in various fields as discussed below:

- Agriculture: The caption for the aerial images will be helpful to the farmers and government for crop identification, crop area determination, crop monitoring and crop damage assessment. The information provided by caption will save human efforts, time and cost of ground surveying as well as reduces chances of errors. It can also be used for soil mapping, soil monitoring and land cover classification, land quality assessment [15]. Eroded soil layers have different colour, tone, structure and hence the process of identification of soil degradation can be automated. Based on soil structure, farmers can be quickly advised by automated system. Information on land cover and changing land cover patterns will be useful to government in planning environment policy and verification. The caption may also help in controlling deforestation by forest cover type and species identification.
- Defence [16-17]: The huge amount of images of nation border, battlefield taken by spy camera, drone or satellite can be captioned and this information can be converted to audio signal and can be sent to soldier or can be used to take immediate necessary actions. The aerial images of sensitive region can be taken periodically. If any unauthorized structure is created in the region then that can be detected as it results in variation in generated caption for the images of same area.
- Visually Impaired Person: Generating image caption would be of great importance for the blind or partially sighted people who cannot access visual information, such as pictures on the internet, in the same way as sighted people can. As visually impaired people can read with the help of Braille script, they can understand an image from its caption printed in Braille script. Alternately generated caption can be converted into audio signal for them.
- Weather Forecasting & Disaster Management [18]: Caption of aerial images can also help in predicting weather conditions and issuing warnings to local residents well in advance in case of disaster such as tsunami.
- Image Retrieval [15-16]: When user enters a caption of required image, search engine will find the image whose caption matches with the user query. This way user image search experience can be enhanced by using this algorithm.
- Social Networking Sites: The user will be assisted by suggesting caption to uploaded photo on social networking sites.
- Scene Classification [15]: The generated captions can also be useful for classifying images into categories which news agency may find useful.

3. CHALLENGES

Though researchers have extensively worked on generating caption for natural images [19-48] generating caption for aerial images is a challenging task as these images exhibits following special characteristics [49]:

- As the aerial images have region with similar characteristics, it is difficult to distinguish sub-regions. If the aerial images contains green plants that may covers green trees, green crops or green grass.
- Natural images are captured by focusing on a specific object. However aerial images are not focusing on specific object as they are taken from top. Hence while generating caption for aerial images, all important things from the image must be described in the caption.
- For a natural image, a building is usually from bottom to up, vehicle's tires are at the bottom and animal or person often stands on ground with their feet. But for aerial images there is no such difference. The orientation such as top or bottom or direction such as left or right cannot be estimated from aerial images.
- Aerial images are taken far away hence area covered is huge and objects in the image are smaller [50].

4. APPROACHES

The approaches followed by researchers for caption generation of aerial images, can be broadly categorized into five categories, namely: template based approach, retrieval based approach, encoder – decoder approach, attention based encoder – decoder approach and combination of retrieval based and encoder – decoder approach.

4.1. Template based approach

In this approach, objects from an aerial image are detected and predefined template is used to generate suitable caption by inserting label of detected objects into it. The performance of the algorithm depends heavily on object detection and its labeling. The generated captions using this approach are simple and have fixed structure. In [16] template based approach is followed for caption generation.

The approach proposed for caption generation in [16] uses Fully Convolutional Network (FCN) technology for object detection. There are two stages of caption generation: first, ground level elements are detected at different levels such as key instance detection, envi-element analysis and landscape analysis by a single FCN

model. Element detection is done at three levels because detection of major elements and identifying relationship between them is crucial in capturing salient aspects of an image in description. The FCN model used is based on VGG-f [51], [52] with some changes. In this model, the filters are transferred from VGG-f, which is trained on ImageNet in a fully supervised way. In second stage, classical template-based approach with linguistic constrains is followed for caption generation. The algorithmic flow of the proposed model is shown in Figure 1.

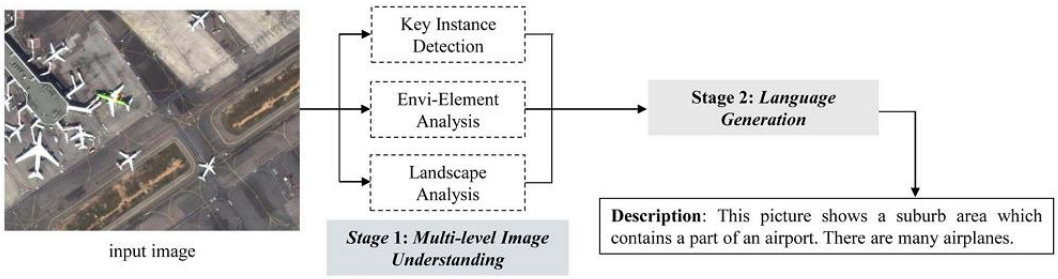


Figure 1. Algorithmic flow proposed in [16]

The Google Earth (RGB) images and GaoFen-2 (GF-2) (multi spectral) satellite images are used for experimentation purpose. The resolution of the Google Earth images is 0.5 m/pixel while that of GF-2 spectral-fused images is 0.8 m/pixel. The images are cropped into small and large sizes with same resolution to compare performance of the algorithm under different scale. The large size image has size ranging from 1000 to 800 pixels while smaller size image has fixed size of 640 X 480 pixels. These images cover ground features of human living environment. But the images do not include regions such as desert or glacier. Only 310 Google Earth large size images are used for training. The 10 large size and 100 small size Google Earth images and 10 large size and 100 small size GF-2 images are used for testing. For objects which are ambiguous due to smaller size or not labelled and are excluded from the data set. Excluding these objects total of 2772 oilpots, 2244 ships, and 1853 airplanes are labeled manually for this experiment.

Instead of objective metrics, subjective criterion is used for evaluation of result as suggested in [25]. Each caption is evaluated by 10 different person and categorized into one of the four level which are "without error", "with minor errors", "related to image" or "unrelated to image" based on quality of generated caption. Though only Google Earth images were used for training, comparable results are obtained for GF-2 images as well. This indicates the strong transfer ability of the proposed approach. For Google dataset, 63% captions were categorized as "without error" while for GF-2 dataset 48% were in "without error" category. Authors suggested integrating contextual information in caption generation process to improve accuracy. They have reported that the quality of the generated caption depends on the correct identification of objects in the first stage. The precision and recall was computed for object detection. Highest precision of 95.3% is reported for oilpot detection while lowest of 84% is for airplane detection. Highest recall of 94.1% is observed for ship detection while lowest of 88.7% is for airplane detection. The confusion matrix of Google Earth images for envi-elements and landscape analysis was presented. Which shows highest accuracy of 99% is obtained for ocean while lowest of 79% for harbor. The authors suggested lack of sufficient samples of harbor during training for poor accuracy. The Graphics Processing Unit (GPU) was used to accelerate FCN model which resulted in faster program execution by 5 to 20 times than a single Central Processing Unit (CPU) thread. The high computational efficiency is achieved due to the use of FCN model which reduces computational redundancy. For larger size images, authors suggested dividing the image into discrete blocks due to graphics memory limitation.

4.2. Retrieval based approach

In this approach, dataset is searched to find image similar to test image and then caption of most identical image from dataset is used to generate caption for test image. If the dataset does not contain image similar to the query image then the generated caption will be embarrassing. This approach fails to generate novel caption. In [53] retrieval based approach is followed for caption generation.

It is difficult to describe an aerial image with only one sentence as it contains multiple objects and associated relationships. Hence to describe an aerial image with multiple sentences, multi sentence captioning task is proposed in [53]. For this purpose, a framework titled Collective Semantic Metric Learning Framework (CSMLF) is proposed which generates 5 sentences for every test image. Architecture of this approach is shown in Figure 2.

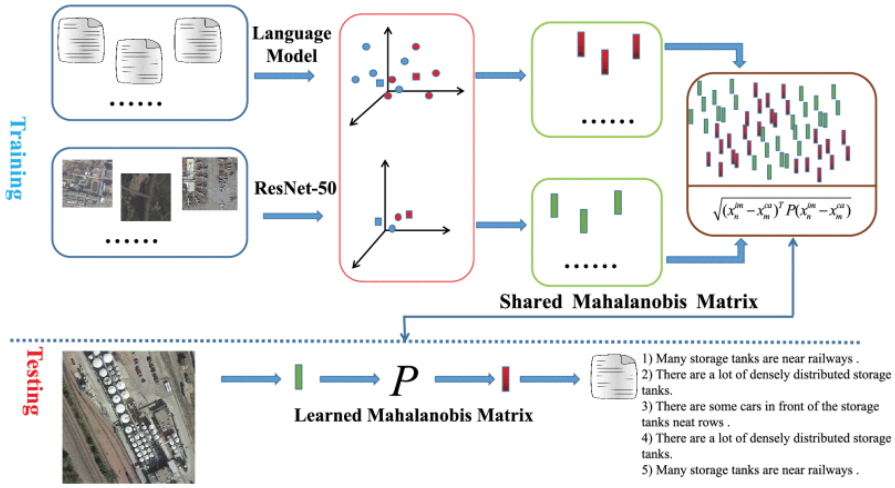


Figure 2. Architecture of multi sentence captioning task proposed in [53]

During training, an aerial image and its corresponding captions are embedded into a common semantic space. The Mahalanobis matrix is learned in the semantic space. An aerial test image is embedded into semantic space and learned Mahalanobis matrix is used to compute distance between test image and collective sentences. Based on the result, closest collective sentences are chosen as the captions for give test image.
Three different methods were experimented for dimension reduction such as Principal Component Analysis (PCA) which is termed as PCSMLF, Canonical Correlation Analysis (CCA) which is termed as CCSMLF and fine-tune procedure which is called as CSMLF (ft). The UCM-Captions, Sydney-Captions and RSICD datasets were used for experimentation. Every dataset was split 80% for training, 10% for evaluation and 10% for test. For performance assessment objective metrics such as BLEU, ROUGE_L, METEOR, CIDEr, and SPICE were used. A subjective criterion for evaluation was also used where captions are categorized into one of the three categories: namely "totally right", "partly right", or "totally wrong" based on quality of generated caption. The 52% captions were categorized as "totally right", 30% as "partly right" on subjective criteria. The results are better for RSICD than other two datasets on all objectives as well as on subjective criteria. On BLEU-1 metric PCSMLF, CCSMLF and CCSMLF (ft) approaches have scored 32.49, 57.59 and 51.06 respectively for RSICD dataset. Still, further research is warranted as the proposed approach failed to create new captions.

4.3. Encoder- decoder approach

In encoder - decoder approach which is referred as generation approach as well, sentence generation is considered to be sequence generation process in continuous way. Most of the approaches uses Convolutional Neural Networks (CNN) for features extraction from image while Recurrent Neural Networks (RNN) or Long Short Term Memory Networks (LSTM) for sentence generation. The LSTM is preferred as it solves the vanishing gradient problem associated with RNN. Encoder – decoder approach for caption generation is followed in [17, 54, 55].

Figure 3 shows architecture of deep multimodal model which is proposed for generating caption for High Spatial Resolution (HSR) remote sensing images in [54]. The model uses CNN to extract visual features from images. These image features and reference sentences are used either by RNN or LSTM to train the model. The trained model is used to generate caption for new images. As it uses both visual and textual information in generating captions it is called as deep multimodal model. Authors have experimented with different CNNs such as AlexNet [56], VGGNet [51] (16-layers net and 19-layers net) and GoogLeNet [51, 56, 67]. They have pre-trained the network on ImageNet dataset and the last full-connected layer output is used as the image feature vector. Then RNN or LSTM is used for caption generation. Experiments were performed on UCM-captions and Sydney-captions dataset. In the experiment, 80% image-captions in the datasets were used as training data, 10% as validation data and the rest 10% as test data. The BLEU, METEOR and CIDEr metrics were used for evaluation of generated captions. It is observed that a result improves from the AlexNet to the VGGNet to the GoogLeNet. Similarly, LSTM gives better results than RNN. Hence combination of VGG-19 with LSTM has given best results considering all evaluation metrics though training LSTM is slower than RNN. For this combination highest score for BLEU-1 metric was obtained as 63.8 and 54.8 on UCM-captions and Sydney-captions dataset respectively. Authors have suggested using bigger dataset to minimize the erroneous results as it will helps the model to distinguish between very similar images.

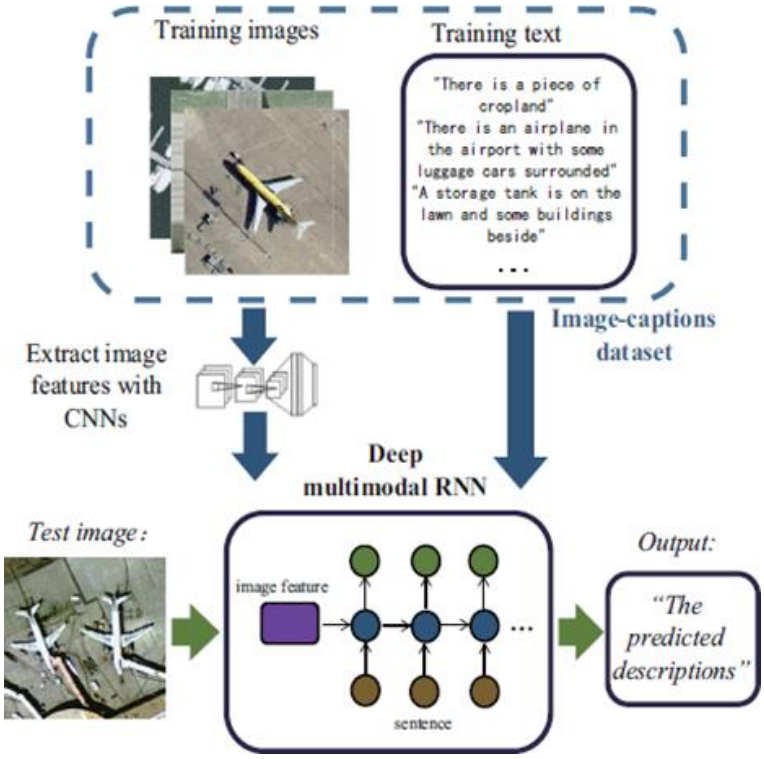


Figure 3. Architecture of deep multimodal model proposed in [54]

For an algorithm proposed in [55], first CNN is used for detection of main objects from aerial image. The CaffeNet is used as pre-trained network with Caffe [56, 58] implementation of CNN. Then result of object detection and its label are used for caption generation. The UCM-captions dataset was used for experimentation purpose. From the dataset, 90% images were used for training purpose while remaining 10% for testing. Subjective criterion was used to evaluate performance of the algorithm. The generated captions were classified into one of the three categories namely: "correct", "partly correct" or "completely incorrect" according to quality. During experimentation, 67% captions were categorized as "correct" while 3% as "completely incorrect". Authors have attributed incorrect results to wrong label prediction in object detection stage. It is also observed that one sentence is not sufficient to describe an image containing large scene as an image contains multiple objects.

Figure 4 shows the framework of multiscale cropping mechanism proposed in [17] which uses CNN and LSTM. A training mechanism is proposed for multi scale cropping which can extract more fine-grained information from images so that generalization performance can be improved and reduces overfitting problem. This cropping mechanism is inspired by ten crops mechanism [56] and multi scale training [59]. For feature extraction from images, CNN is used while for sentence generation LSTM is used. Experimentation was done on UCM-captions and Sydney-captions dataset. To evaluate the performance BLEU metric was computed. In encoding stage, for training and testing three CNNs namely, VGG-16, Inception-ResNetV2 and ResNet152 [60] were used. For sentence generation, LSTM was used with Stochastic Gradient Descent (SDG) for optimization. The ResNet-152 has shown better performance than other CNNs on both the dataset. The ResNet-152 has achieved 59.4 and 61.5 BLEU-1 score for UCM-captions and Sydney-captions dataset respectively.

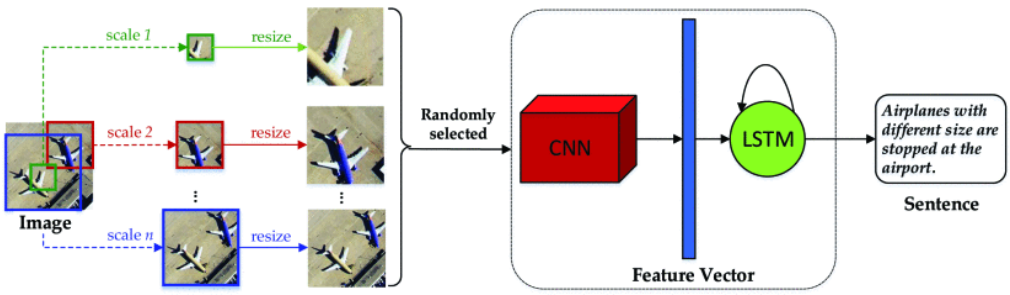


Figure 4. Framework of multi scale cropping mechanism proposed in [17]

4.4. Attention based encoder – decoder approach

Attention mechanism helps the encoder – decoder algorithm in concentrating on important parts of an image so the generated caption will be capturing main aspect of the image. This approach for caption generation is followed in [15, 49, 50, 61-64].

To avoid information loss while captioning an aerial image due to higher pixels and smaller target size, Intensive Positioning Network (IPN) is proposed in [61]. The algorithmic view of this model is shown in Figure 5. The IPN is based on neural network and attention mechanism and can predict regions containing important information. It produces multiple region description blocks around these regions. The proposed algorithm is divided into two steps: first, IPN model outputs area characteristics and regional locations. Then caption is generated by LSTM language model using regional features and locations. Google Earth images and GF-2 satellite images with label dataset which were used in [16], are used for experimentation with similar training and testing image ratio. Subjective criterion introduced in [16] was used to check quality of generated captions. For Google earth dataset, 80% and 63% captions were categorized as "without errors" for proposed approach and approach used in [16] respectively. For GF-2 dataset, 73% and 48% captions were categorized as "without errors" for proposed approach and approach used in [16] respectively. For 480 X 600 image, the proposed approach and approach used in [16] takes 1.11 and 1.26 seconds execution time respectively. For 1600 X 2400 image, the proposed approach and approach used in [16] takes 3.70 and 5.84 seconds execution time respectively. Hence the performance of the proposed algorithm is better compared with [16] on the basis of quality of generated caption as well as on execution time parameter.

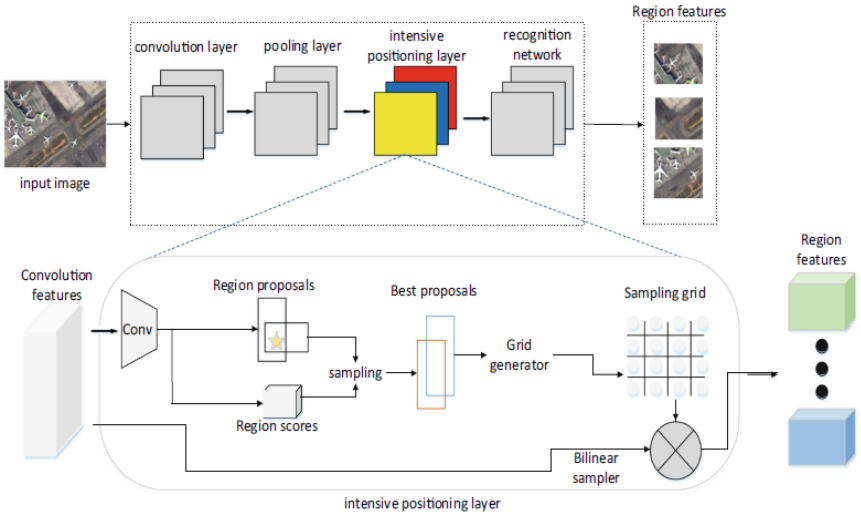


Figure 5. Algorithmic view of model proposed in [61]

Special characteristics of aerial images such as scale ambiguity, category ambiguity, and rotation ambiguity are considered for caption generation in [49]. The proposed approach is based on encoder – decoder framework [27] and shown in Figure 6. Here an image is encoded into a vector and then vector is decoded into a sentence. Authors have experimented with multimodal and annotation based methods for caption generation task.

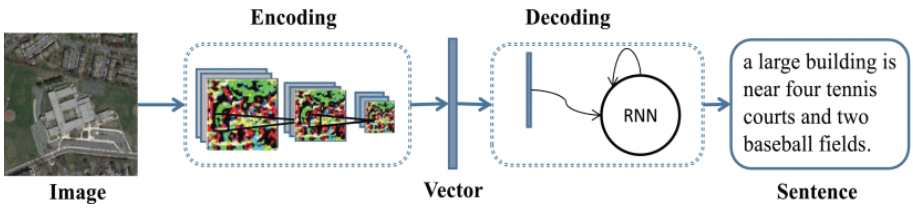


Figure 6. Outline of encoder–decoder model proposed in [49]

The multimodal method is based on [54] and utilizes a deep multimodal neural network for caption generation. The caption generation process involves representing an image, representing sentence and sentence generation. To represent aerial images, two different methods were used: handcrafted feature method and learned feature method. In the first method, handcrafted features are extracted from image and then any of the feature encoding techniques such as Scale-Invariant Feature Transform (SIFT) [65], Bag Of Words (BOW)

[66], Fisher Vector (FV) [67], or Vector of Locally Aggregated Descriptors (VLAD) [68] is used to obtain image representation.

The learned feature method uses machine learning algorithm to automatically learn image representation from training data. For feature extraction of aerial images, fully connected layers of several CNNs including AlexNet [56], VGGNet [51], and GoogLeNet [57], pre-trained on ImageNet dataset is used. For sentence generation, RNN or LSTM is used in multimodal method.

The attention-based method is based on [25]. This method uses either deterministic manner (soft attention) or stochastic manner (hard attention) to train the model. The former uses standard back propagation technique while later trains the model by maximizing a lower bound. For sentence generation LSTM is used. For experimentation, UCM-Captions, Sydney-Captions and RSICD dataset were used. To assess quality of generated captions objective metrics such as BLEU, ROUGE_L, METEOR, and CIDEr were used along with subjective criterion. For multimodal method, 80% samples were used for training, 10% for validation, and 10% for testing. Among all four handcrafted representations, VLAD performs the best on UCM-captions and RSICD dataset. The VLAD-LSTM combination has achieved 70.15 and 50.03 BLEU-1 score for UCM-captions and RSICD datasets respectively. The FV have produced better result on Sydney-captions dataset. The FV-LSTM combination has achieved 63.31 BLEU-1 score for Sydney-captions dataset. It is reported that the results for LSTM were better for all the data set. For RSICD dataset, BLEU score for VLAD-RNN is 49.37 compared to 50.03 for VLAD-LSTM combination.

In case of learned feature method, authors have also compared performance of different CNNs such as VGG-19, VGG-16, AlexNet, GoogLeNet on RSICD dataset using LSTM. Though much difference is not observed in performance, VGG-19 had provided better result on BLEU (BLEU-1 score is 58.33) and METEOR (Score is 26.13) metrics while Alexnet had provided better results on ROUGE_L (Score is 51.91) and CIDEr (Score is 41.05) metrics than others. Overall, CNNs had shown good results compared to handcrafted representation.

For experimentation of attention-based method, different CNNs such as VGG-19, VGG-16, AlexNet, GoogLeNet were used with LSTM. The hard attention mechanism based on GoogLeNet had provided better result for UCM-captions data set (BLEU-1 score is 83.75) and RSICD data set (BLEU-1 score is 68.81). The soft attention mechanism based on VGG-16 had provided better results for Sydney-captions dataset (BLEU-1 score is 73.21).

For few images, the generated caption contains objects which were not present in the images. Authors have attributed this to high frequency co-occurrence of two words in the dataset. So if one object is present in an image then word associated with commonly co-occurred object is also present in the caption. To verify generalization capability, models were trained on dataset which is different than the one used for testing. The subjective criterion was used for evaluation. Here, captions were classified into three categories namely: “totally depict image”, “related to image” or “unrelated to image” based on quality. The model trained on RSICD can generate 38% and 40 % captions relevant to the test images from UCM-captions and Sydney-caption dataset respectively. Hence, generalization capability of model trained on RSICD is better than model trained on UCM-captions or Sydney-captions dataset.

Reference [15] follows region driven approach which uses domain probabilities for caption generation. It follows encoding – decoding approach and shown in Figure 7. During encoding, features are extracted from images using CNN and caption sequence is processed using text processing. During decoding, the encoded features are decoded for caption generation using attention-based bidirectional LSTM.

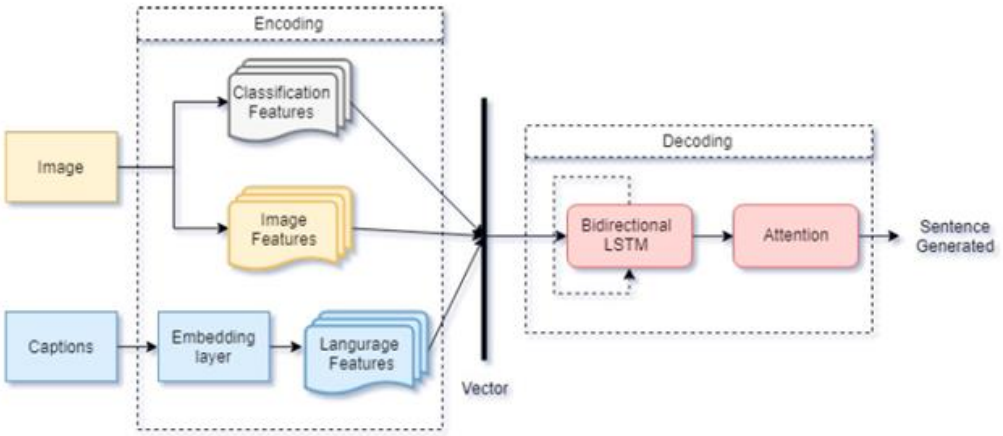


Figure 7. Framework of the Region Driven Remote Sensing Image captioning approach proposed in [15]

Experimentation was done on UCM-captions and UAVIC dataset. Each dataset was divided as 80% for training, 10% for testing and 10% for validation. For implementation of algorithm, Keras in anaconda with Tensorflow was used. Google Colab was used to run the model and pre-trained word2vec [69] model was used for text processing. Performance of algorithm was evaluated using BLEU metric. Result shows that addition of LSTM layer resulted in improved accuracy. Use of ResNet152 for feature extraction had provided comparatively better results than VGG-19, ResNet50 & InceptionV3. Use of bidirectional LSTM with attention resulted in better accuracy and generation of coherent sentences. The proposed approach has achieved BLEU-1 score of 84 for UCM-captions dataset.

The model proposed in [62], follows encoder – decoder framework with attributes attention mechanism. The structure of this model is shown in Figure 8. Use of this mechanism results in paying attentions to the relation between input image and generated words. Here, VGG-16 CNN is used for features and attributes extraction and LSTM is used for caption generation. The attributes attention mechanism helps in generating different words for different regions of images. It is also useful in paying attention to high level features along with low level features of an image.

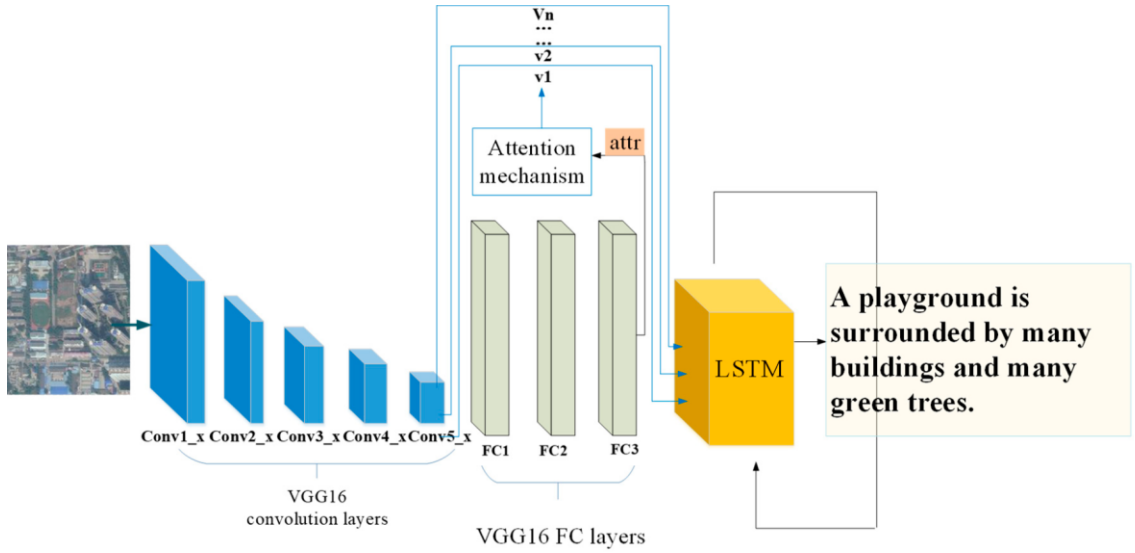


Figure 8. The structure of model proposed in [62], attr can be the output of the last fully connected layer or softmax layer

Experiments were performed on UCM-Captions, Sydney-Captions and RSICD dataset. For Sydney-captions dataset, 81% images were used for training while for UCM-captions and RSICD datasets 80% images were used for training, while remaining images were used for validation and testing. Objective metrics such as BLEU, ROUGE_L, METEOR, and CIDEr were used for evaluation of generated sentences. The performance of proposed approach is compared with the approaches proposed in [49, 53, 54]. The proposed approach has two variants namely FC-Att+LSTM and SM-Att+LSTM. The former variant uses output of last fully connected layer of VGG-16 as attributes to affect attention mechanism while the later uses output of softmax layer of VGG-16 as attributes. For BLEU-1 metric, score achieved by SM-Att+LSTM, FC-Att+LSTM, [49, 53, 54] are 81.43, 80.76, 79.05, 59.98 and 69.66 respectively for Sydney-captions dataset. The result shows that the proposed approach has provided better results than other approaches on all three datasets. The performance of SM-Att+LSTM is superior to that of FC-Att+LSTM. There are few examples of generation of wrong captions as well. The caption describes the object which is not present in the image. Authors attributed this to high frequency of some words appearing together during training. There are few examples of misrecognition as well which is resulted due to small interclass dissimilarity.

The earlier attention mechanism based image captioning algorithms were based on passive attention. The passive attention mechanism focuses on ground truth caption while active attention mechanism focuses on human observation of the image. The sound active attention framework proposed in [63], uses human sound as guiding information for active attention. The framework is shown in Figure 9 and consists of four modules namely: representation module, sound Gated Recurrent Unit (GRU) module, attention GRU module and output GRU module.

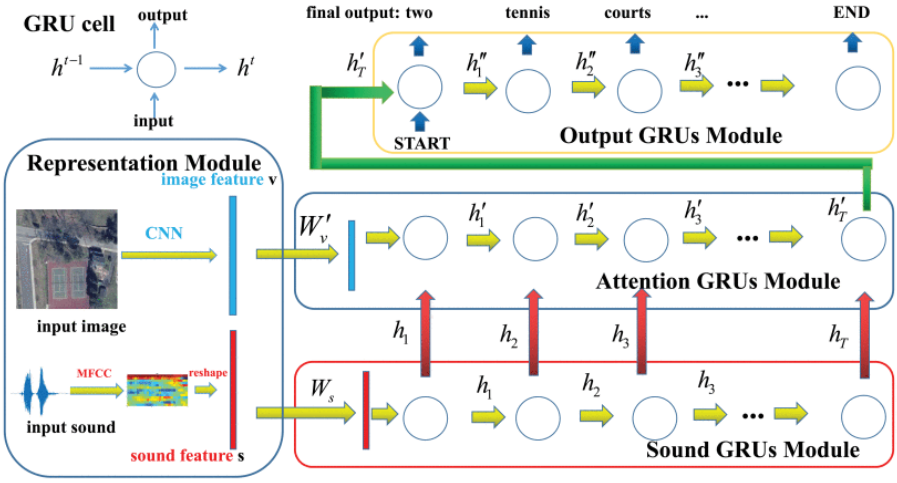


Figure 9. Framework of model proposed in [63]

In representation module, sound features are extracted by Mel – Frequency Cepstral Coefficient (MFCC) and CNN is used to extract features of an input image. Then sound features are imported into sound GRU module. Its output guides attention GRU module which also uses extracted image features. Output GRU module generates caption by importing output of attention GRU module.

For experimentation purpose, UCM–captions, Sydney-captions and RSICD datasets were used. To apply proposed algorithm, each aerial image must be associated with human sound consisting one word describing an image. But as the available datasets does not contain such sound information for an aerial image, sound dataset was prepared in [63]. Guided sounds for the datasets were generated using sound generation application and it is referred as "fake dataset". Similarly "real sound dataset" was developed only for UCM–captions dataset by manually recording human sound. Every dataset was split as 80% for training, 10% for evaluation and rest for test. The algorithm was implemented using Keras deep learning framework. For performance assessment, objective metrics such as BLEU, ROUGE_L, METEOR, CIDEr and SPICE were used. The BLEU-1 score for proposed approach, and for [49, 54] is 83.26, 81.57 and 63.50 respectively when experimented on UCM-captions dataset. The performance of proposed algorithm was better when compared with [49, 54]. The SPICE score for proposed approach on UCM-captions "fake sound" and UCM-captions "real sound" dataset is 47.87 and 39.53 respectively. Authors have attributed to noise effect of real sound for the lower performance of real sound dataset compared with fake sound dataset. A subjective criterion was also used to evaluate generated captions. The captions are divided into three categories namely: "totally right", "partly right" or "totally wrong" according to the quality. The results for UCM-captions (77% "totally right") and Sydney dataset (76% "totally right") found to be much better than that of RSICD dataset (40% "totally right").

The end to end captioning algorithm with scene attention mechanism for aerial images is proposed in [50] and outlined in Figure 10. The CNN is used to extract features from an input image. Then proposed scene attention mechanism generates context vector using extracted features and LSTM. Finally, the caption is generated by using LSTM and context vector. Scene attention mechanism focuses on important part of the images by using semantic information from LSTM and global visual information from features. This helps in caption generation.

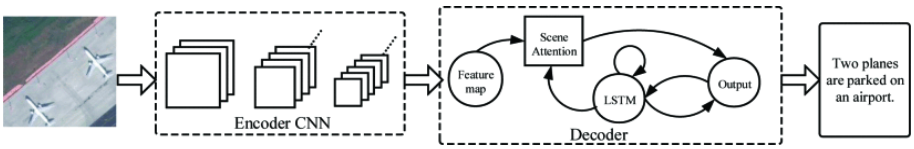


Figure 10. The outline of the model method in [50]

The UCM-Captions, Sydney-Captions and RSICD datasets were used for experimentation. Every dataset was split as 10% for validation, 10 % for test and rest for training purpose. VGG-19 network pre-trained on ImageNet dataset was used as an encoder [51]. For performance assessment, objective metrics such as BLEU, ROUGE_L, METEOR, CIDEr were used. The BLEU-1 score for proposed approach, and for [49, 53, 54] is 78.60, 74.40, 44.40, and 71.50 respectively when experimented on Sydney-captions dataset. Performance of proposed algorithm was compared with [49, 53, 54] and found to be better.

To solve overfitting problem and to use semantic information, Variational Autoencoder and Reinforcement Learning based Two Stage Multi Task Learning Model (VRTMM) is proposed in [64]. The structure of proposed VRTMM is depicted in Figure 11. Firstly, CNN is fine-tuned on dataset jointly with variational autoencoder, later caption is generated by transformer using spatial and semantic features. Quality of generated caption is improved by applying reinforcement learning. For fine-tuning purpose, RESISC45 [70] dataset was used while for experimentation RSICD dataset was used. The objective metrics such as BLEU, ROUGE_L, METEOR, and CIDEr were used to evaluate performance. The BLEU-1 score for proposed approach, and for [49, 53, 54, 62] on RSICD dataset is 79.34, 73.36, 57.59, 63.78, and 75.71 respectively. Comparison of proposed approach with [49, 53, 54, 62] shown its superiority on all objective metrics.

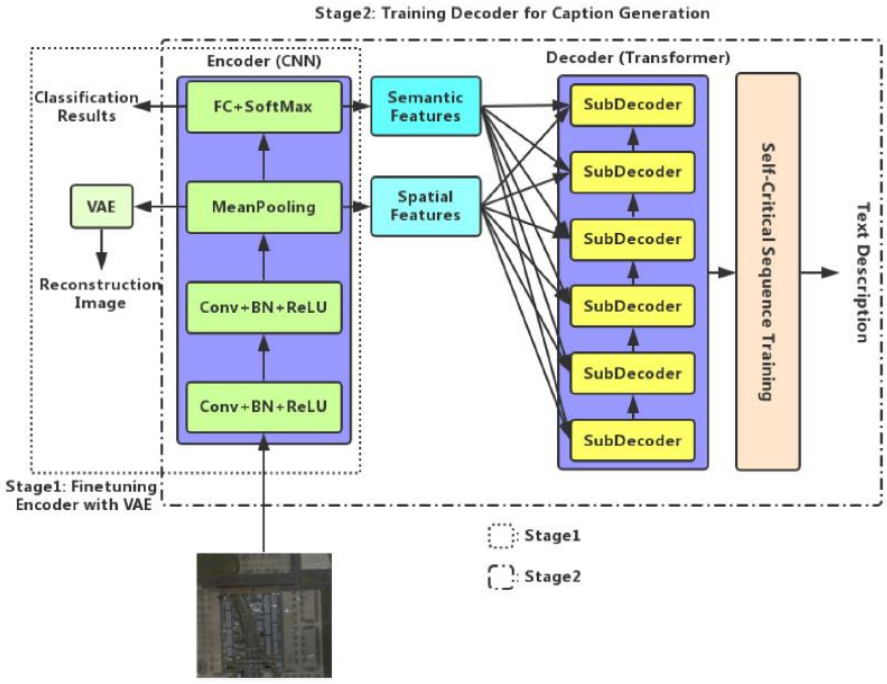


Figure 11. The structure of VRTMM proposed in [64], encoder is a pre-trained convolutional neural network such as VGG-16 while decoder is the transformer

4.5. Combination of retrieval based and encoder – decoder approach

Retrieval based approach of caption generation fails to generate novel caption whereas encoder – decoder approach may generate irrelevant caption for the given image. Hence researchers have experimented with combination of these approaches in [18, 71].

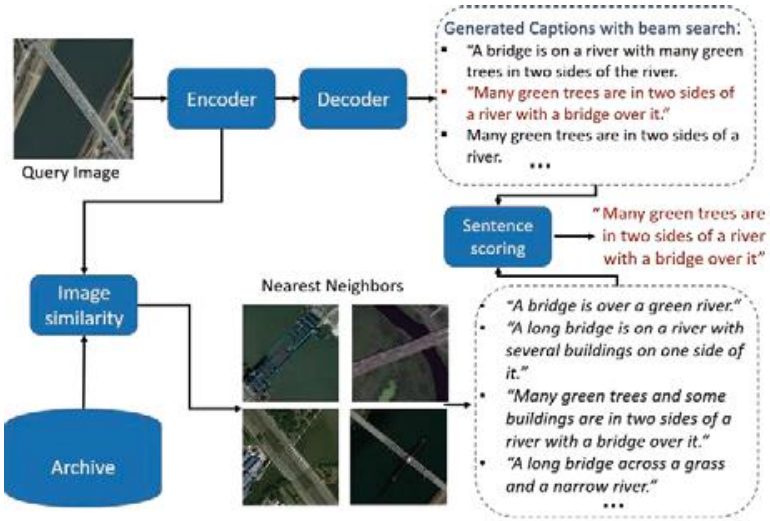


Figure 12. Scheme block of proposed system in [71]

Reference [71] proposes combination of retrieval based approach and encoder – decoder approach for image captioning which is depicted in Figure 12. First, multiple captions are generated using beam-search algorithm then best caption is chosen on the basis of lexical similarity with the caption of retrieved image. The lexical similarity is measured using consensus caption score [72].

The proposed encoder - decoder approach is similar to [25], it uses InceptionV3 CNN for feature extraction and RNN for caption generation. Beam-search algorithm is used to generate multiple captions. Nearest neighbor approach is followed to find similar image from archive and then caption consensus score is computed to find lexical similarity. The caption having highest lexical similarity is chosen based on this consensus score. The lexical similarity is calculated using BLEU score. The RSICD dataset was used for experimentation and BLEU score is computed for performance evaluation. On BLEU-1 metric, the proposed approach achieved 66.1 score while simple encoder-decoder approach achieved 65.7 score. The results have shown that the proposed approach is superior to simple encoder-decoder approach.

The Retrieval Topic Recurrent Memory Network (RTRMN) is proposed in [18], uses topic words to generate caption for aerial images. During training process, topic repository is built of topic words extracted from caption dataset. The topic word contains common determinate information obtained from various captions of same training image. To extract topic words from caption dataset, either semantic topics or statistical topics method is used.

For test image, image features are extracted by ResNet-101 [60]. Topic words of an image similar to test image are obtained from topic repository by random sample retrieval. Using these topic words and manually edited topic words caption is generated by Recurrent Memory Network (RMN) with the help of CNN. The Euclidian Distance was selected as distance measure method to get the topic information for the test image. The algorithm was implemented using Keras. Figure 13 shows the architecture of RTRMN proposed in [18].

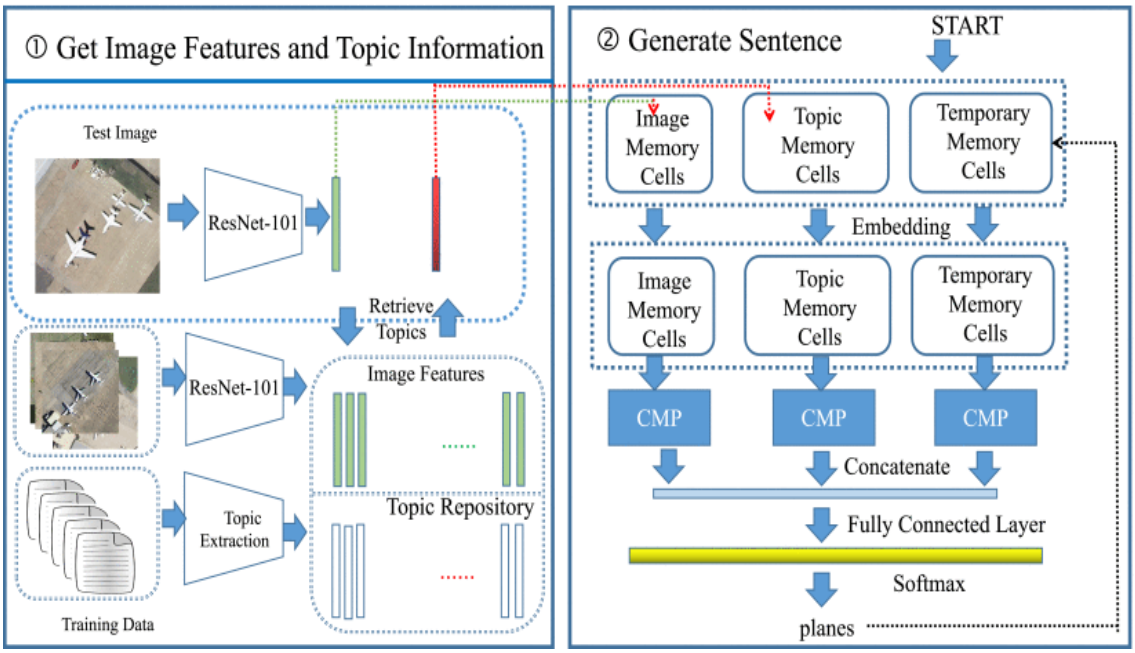


Figure 13. RTRMN architecture proposed in [18]

The experiments were performed on UCM-Captions and RSICD dataset. To evaluate performance of an algorithm, objective metrics such as BLEU, ROUGE_L, METEOR, CIDEr and SPICE were used. Better performance was observed for UCM-captions than RSICD caption dataset. On BLEU-1 metric, [53, 73, 74] achieves 36.71, 40.59 and 36.97 score respectively while the proposed algorithm achieves 80.28 score for UCM-captions dataset. Hence performance of the proposed algorithm is better than [53, 73, 74]. To avoid irreverent words and redundant information in generated caption, edition of topic word which is termed as controllability of caption generation is suggested.

Table 1 summarizes different approaches followed for aerial image caption generation with information about datasets and evaluation measures used.

Table 1. An overview of the approaches, datasets, and evaluation measures reviewed in this survey

Reference	Approach	Datasets	Evaluation Measures
[16]	Template based	Google Earth images and GaoFen-2 (GF-2) satellite images with label	Subjective criterion
[53]	Retrieval based	UCM-Captions, Sydney-Captions, RSICD	BLEU, ROUGE_L, METEOR, CIDEr, SPICE and subjective criterion
[54]	Encoder – decoder	UCM-captions, Sydney-captions	BLEU, METEOR, CIDEr
[55]	Encoder – decoder	UCM-captions	Subjective criterion
[17]	Encoder - decoder	UCM-Captions, Sydney-Captions	BLEU
[61]	Attention based encoder – decoder	Google Earth images and GaoFen-2 (GF-2) satellite images with label	Subjective criterion
[49]	Attention based encoder – decoder	UCM-Captions, Sydney-Captions, RSICD	BLEU, ROUGE_L, METEOR, CIDEr and subjective criterion
[15]	Attention based encoder – decoder	UCM-Captions, UAVIC-Captions	BLEU
[62]	Attention based encoder – decoder	UCM-Captions, Sydney-Captions, RSICD	BLEU, ROUGE_L, METEOR, CIDEr
[63]	Attention based encoder – decoder	UCM-Captions, Sydney-Captions, RSICD, UCM-Captions with fake sound and real sound dataset	BLEU, ROUGE_L, METEOR, CIDEr, SPICE and subjective criterion
[50]	Attention based encoder – decoder	UCM-Captions, Sydney-Captions, RSICD	BLEU, ROUGE_L, METEOR, CIDEr
[64]	Attention based encoder – decoder	RSICD	BLEU, ROUGE_L, METEOR, CIDEr
[71]	Combination of retrieval based and encoder – decoder approach	RSICD	BLEU
[18]	Combination of retrieval based and encoder – decoder approach	UCM-Captions, RSICD	BLEU, ROUGE_L, METEOR, CIDEr, SPICE

5. DATASETS

Following datasets are available and used by different researchers in their experimentation.

5.1. UCM Captions Dataset

The UC Merced (UCM) dataset is proposed in [54]. This dataset is based on UC Merced Land Use Dataset [75]. The aerial images are classified into 21 categories where in each category there are 100 images. So, the dataset contains 2100 images of size 256 x 256 pixels. The image categories are agricultural, airplane, baseball diamond, beach buildings, chaparral, dense residential, forest, freeway, golf course, harbor, intersection, medium residential, mobile home park, overpass, parking lot, river, runway, sparse residential, storage tanks, tennis court. The pixel resolution of these images is 0.3048m. The images were extracted manually from the USGS National Map Urban Area Imagery collection [75]. Each image in described with five different captions hence there are 10500 captions in the dataset. Figure 14 provides an image and associated captions from UCM-Captions dataset.

1. *An intersection with some cars on the road.*
2. *An intersection with some houses and plants at the corners.*
3. *An intersection with some houses and plants at the corners.*
4. *An intersection with two roads vertical to each other.*
5. *An intersection with two roads vertical to each other.*

Figure 14. Example of an image and corresponding five captions from UCM-Captions dataset

5.2. Sydney Captions Dataset

The Sydney captions dataset is also proposed in [54]. This dataset is based on Sydney dataset [76]. The aerial images are classified into 7 categories and there are total 613 images in the dataset. The categories are residential, airport, meadow, rivers, ocean, industrial, and runway and size of every image is 500×500 pixels [49]. The pixel resolution of the image is 0.5 m. These images are of Sydney city, Australia, taken from Google Earth. Each image in described with five different captions hence there are 3065 captions in the dataset. Figure 15 provides an image and associated captions from Sydney-Captions dataset.

1. *Lots of houses with red and orange roofs arranged in lines.*
2. *A residential area with houses arranged neatly and some roads go through this area.*
3. *A town with many houses arranged neatly and divided by some roads.*
4. *A residential area with houses arranged neatly while many plants on the roadside.*
5. *A residential area with houses densely arranged and some crossroads in the middle.*

Figure 15. Example of an image and corresponding five captions from Sydney-Caption Dataset

5.3. Remote Sensing Image Captioning Dataset (RSICD)

The RSICD dataset is proposed in [49]. The captions are provided by experienced volunteers. Following instructions for caption generation were provided to volunteers based on [73, 77, 78].

- To describe all the major parts of an image.
- When there are multiple objects present in an image, do not start sentence "There is".
- Avoid use of vague or generalized words like large, tall, and many in the absence of contrast.
- Avoid use of direction nouns, such as north, south, east, and west.
- The caption must contain six words.

There are 10,921 images in dataset classified into 30 classes. The classes are airport, bare land, baseball field, beach, bridge, center, church, commercial, dense residential, desert, farmland, forest, industrial, meadow, medium residential, mountain, park, school, square, parking, playground, pond, viaduct, port, railway station, resort, river, sparse residential, storage tanks, and stadium. These images are collected from Google Earth [79], Baidu Map, MapABC, and Tianditu. The images are having size 224 × 224 pixels with various resolutions. The 724 images are described using 5 different sentences, 1495 images are described using 4 different sentences, 2182 images are described using 3 different sentences, 1667 images are described using 2 different sentences while remaining 4853 sentences are described using only 1 sentence. Each image in the dataset is described using 5 sentences. So, existing sentences are duplicated for images who were not described by 5 sentences. Hence, there are total 54,605 sentences for 10,921 images in the dataset. Figure 16 provides an image and associated captions from RSICD dataset.

1. *Four planes are stopped on the open space between the parking lot.*
2. *Four white planes are between two white buildings.*
3. *Some cars and two buildings are near four planes.*
4. *Four planes are parked next to two buildings on an airport.*
5. *Four white planes are between two white buildings.*

Figure 16. Example of an image and corresponding five captions from RSICD Dataset

5.4. UAVIC Dataset

The UAVIC dataset was proposed in [15], which is based on images and videos collected from Team Dhaksha [80] of MIT. The images are classified into 12 classes namely, barren lands, farm lands, forests,

gardens, highways, playgrounds, residential, roads, runway, solar panels, water bodies, temple. The shape of the images is 400 X 400 pixels. Each image of the dataset is described using 5 captions.

6. EVALUATION MEASURES

To evaluate the quality of caption generated by algorithm, following ways are followed by researchers: objective metrics and subjective criteria.

6.1. Objective Metrics

An automatic evaluation of caption for given image is quick, inexpensive, language independent therefore widely used by the researchers. This is done by comparing generated caption with reference captions from the dataset. Following are commonly used objective metrics for evaluation of caption generated for aerial images. For all objective metrics, higher metric score indicates better match of generated caption with the reference caption.

6.1.1. BLEU (BiLingual Evaluation Understudy)

The BLEU metric is proposed in [81] and uses weighted average of variable length phrase matches against the reference sentence. It measures the co-occurrence of n-gram between the generated caption and the reference caption from the dataset where n-gram is set of one or more ordered words. The BLEU is precision based score which ranges from 0 to 1 where higher score indicates better match.
The brevity penalty (BP) can be defined as,

$$BP = \begin{cases} 1, & if\ c > r \\ e^{\left(1-\frac{r}{c}\right)}, & if\ c \ll r \end{cases} \quad (1)$$

where, c is the length of the generated caption and r is the reference caption length.

The BLEU score can be calculated as,

$$BLEU = BP \cdot exp\left(\sum_{n=1}^{N} w_n \log p_n\right) \quad (2)$$

where, p_n is modified n-gram precision, N is n-gram length, w_n is the positive weights and the sum of w_n is one.

6.1.2. ROUGE (Recall Oriented Understudy for Gisting Evaluation)

The ROUGE metric is proposed in [82] and it counts n-gram, word sequences and word pairs between generated caption and reference captions from the dataset. It is recall based measure and different variants such as ROUGE-N, ROUGE-L, ROUGE-W and ROUGE-S are available for use. ROUGE_L calculates F-measures between X of m length and Y of n length for given Longest Common Subsequence (LCS) where, X is reference caption while Y is generated caption.

$$R_{lcs} = \frac{LCS\ (X,Y)}{m} \quad (3)$$

$$P_{lcs} = \frac{LCS\ (X,Y)}{n} \quad (4)$$

$$\beta = \frac{P_{lcs}}{R_{lcs}} \quad (5)$$

The ROUGE score can be calculated as,

$$ROUGE\_L = \frac{(1+\beta^2)R_{lcs}P_{lcs}}{R_{lcs}+\beta^2 P_{lcs}} \quad (6)$$

where, LCS (X, Y) is the LCS of X and Y.

6.1.3. METEOR (Metric for Evaluation of Translation with Explicit ORdering) Universal

The METEOR Universal score was proposed in [83] and it is a version of METEOR metric. It supports language specific evaluations. The score is computed by first aligning the generated caption with reference caption and then by calculating sentence – level similarity.

F_{mean} is calculated based on unigram precision (P) and unigram recall (R) as,

$$F_{mean} = \frac{10\,P\,R}{R + 9\,P} \tag{7}$$

Penalty is computed as,

$$Penalty = 0.5\left(\frac{chunks_{num}}{unigrams\_matched_{num}}\right)^3 \tag{8}$$

where, the numerator $chunks_{num}$ is the number of chunks and the denominator $unigrams\_matched_{num}$ is the number of matched unigrams.

The METEOR is calculated as,

$$METEOR = F_{mean}\,(1 - Penalty) \tag{9}$$

6.1.4. CIDEr (Census based Image Description Evaluation)

The CIDEr metric is proposed in [84]. It measures the similarity of generated caption with multiple reference captions. The CIDEr matric shows high agreement with consensus as evaluated by humans. A version CIDEr-D is also available for systematic evaluation and benchmarking. The Term Frequency - Inverse Document Frequency (TF-IDF) weighting $g_k()$ for every n-gram is computed to find CIDEr.

$$CIDEr_n(Y, X) = \frac{1}{m}\sum_{j=1}^{m}\frac{g_n(Y)\cdot g_n(X_j)}{\|g_n(Y)\|\|g_n(X_j)\|} \tag{10}$$

where, X_j is the j^{th} reference caption, Y is the generated caption and m is number of reference captions.

CIDEr is a weighting sum of all the $CIDEr_n$

$$CIDEr(Y, X) = \sum_{n=1}^{N} w_n CIDEr_n(Y, X) \tag{11}$$

where, uniform weights w_n=1/N

6.1.5. SPICE (Semantic Propositional Image Caption Evaluation)

The SPICE metric is proposed in [85]. The BLEU, ROUGE, CIDEr and METEOR metrics are sensitive to n-gram overlap unnecessarily. This drawback is overcome in SPICE. It evaluates the caption by transforming it into graph based semantic representation. The SPICE has shown better performance compared to BLEU, ROUGE, CIDEr and METEOR metrics in terms of agreement with human evaluation.

6.2. Subjective Criteria

The objective metrics compares the generated caption with the reference captions from the dataset and computes the score. Every individual has different style and way of describing an image and a set of reference captions from the dataset cannot captures all such styles and ways. The comparison of generated caption with limited set of reference caption may also results in bias hence in [16, 49, 53, 55, 61, 63] generated captions were evaluated by humans into various categories based on appropriateness of caption to the corresponding image.

The approach followed in [16, 61] is proposed in [25]. Here the generated caption is categorized based on its appropriateness and relevance to the corresponding image, into four categories namely: “without errors,” “with minor errors,” “related to the image,” or “unrelated to the image” by the humans. The “without errors” category indicates caption depicts the image correctly while “with minor errors” category indicates that captions depicts the image but with minor errors. The “related to the image” category indicates that caption fails to depict the image correctly but it is related to image while last category indicates caption is completely unrelated to the image.

In [55], humans have categories generated caption into one of three categories namely: “correct”, “partly correct” or “completely incorrect”, while in [49], captions are categorized into one of three categories namely: “totally depict image”, “related to image” or “unrelated to image”. In [53, 63] humans have categories generated caption into one of three categories namely: “totally right”, “partly right” or “totally wrong”.

7. FUTURE DIRECTIONS

Though remarkable progress is achieved in caption generation task for an aerial image, the generated caption fall short of the human performance. The caption generated by existing approaches are simple and

describe only the objects and their relationships. To exploit the applications of aerial image caption generation task, the generated caption must provide in-depth and specific information of an image. Describing an image using a single sentence may not cover all aspects of an image. Hence focus can be given on describing an image using multiple sentences in future.

In some cases, generated caption mistakenly describes an object which is not present in an image. This happens due to frequent co-occurrence of two words in the dataset. Hence, for a word corresponding to any of the object present in the image, co-occurring word is also included in the generated caption. Creating comprehensive and diverse dataset with huge number of images and sufficient number of corresponding captions will help to solve this issue. The dataset should contains at least 5 unique captions for each image to improve the vocabulary of the trained model.

The attention based encoder – decoder approach has provided significantly better quality caption than other approaches on both objective metrics and subjective criteria. However, there are examples of generation of irrelevant and erroneous captions. Consideration of both high and low level features of an image during feature extraction stage at encoder is key to this solve this issue.

Use of subjective criteria to evaluate the quality of generated caption is highly recommended as none of the objective metrics score truly reflect the quality of generated caption.

8. CONCLUSIONS

Due to special nature of aerial images, generating a caption for them is a complex task. In this survey, the approaches followed for aerial image caption generation are broadly categorized into five categories namely: template based approach, retrieval based approach, encoder – decoder approach, attention based encoder – decoder approach and combination of retrieval based and encoder – decoder approach and discussed comprehensively. The UCM-captions, Sydney-captions, RSICD and UAVIC datasets used for experimentation are described in details. The objective metrics such as BLEU, ROUGE, METEOR, CIDEr, SPICE and subjective criteria used for evaluation of generated caption is explored. Future directions are provided to fully exploit the benefits of aerial image caption generation algorithms, and to improve the quality of generated captions.

REFERENCES

[1] O. A. B. Penatti, K. Nogueira, and J. A. dos Santos, "Do deep features generalize from everyday objects to remote sensing and aerial scenes domains?" in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops, 2015, pp. 44–51.

[2] J. Inglada, "Automatic recognition of man-made objects in high resolution optical remote sensing images by svm classification of geometric image features," ISPRS J. Photogramm. Remote Sens., vol. 62, no. 3, pp. 236–248, 2007.

[3] Z. Zou and Z. Shi, "Ship detection in spaceborne optical image with SVD networks," IEEE Trans. Geosci. Remote Sens., vol. 54, no. 10, pp. 5832–5845, Oct. 2016.

[4] L. Zhang, Z. Shi, and J. Wu, "A hierarchical oil tank detector with deep surrounding features for high-resolution optical satellite imagery," IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens., vol. 8, no. 10, pp. 4895–4909, Oct. 2015.

[5] Z. An, Z. Shi, X. Teng, X. Yu, and W. Tang, "An automated airplane detection system for large panchromatic image with high spatial resolution," Optik-Int. J. Light Electron Opt., vol. 125, no. 12, pp. 2768–2775, Jun. 2014.

[6] J. Han, D. Zhang, G. Cheng, L. Guo, and J. Ren, "Object detection in optical remote sensing images based on weakly supervised learning and high-level feature learning," IEEE Trans. Geosci. Remote Sens., vol. 53, no. 6, pp. 3325–3337, Jun. 2015.

[7] Z. Shi, X. Yu, Z. Jiang, and B. Li, "Ship detection in high-resolution optical imagery based on anomaly detector and local shape feature," IEEE Trans. Geosci. Remote Sens., vol. 52, no. 8, pp. 4511–4523, Aug. 2014.

[8] V. Risojevi´c and Z. Babi´c, "Unsupervised quaternion feature learning for remote sensing image classification," IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens., vol. 9, no. 4, pp. 1521–1531, Apr. 2016.

[9] B. Demir and L. Bruzzone, "Hashing-based scalable remote sensing image search and retrieval in large archives," IEEE Trans. Geosci.Remote Sens., vol. 54, no. 2, pp. 892–904, Feb. 2016.

[10] X. Lu, X. Zheng, and Y. Yuan, "Remote sensing scene classification by unsupervised representation learning," IEEE Trans. Geosci. Remote Sens., vol. 55, no. 9, pp. 5148–5157, Sep. 2017.

[11] G. Cheng, J. Han, L. Guo, Z. Liu, S. Bu, and J. Ren, "Effective and efficient midlevel visual elements-oriented land-use classification using VHR remote sensing images," IEEE Trans. Geosci. Remote Sens., vol. 53, no. 8, pp. 4238–4249, Aug. 2015.

[12] D. Tuia, F. Ratle, F. Pacifici, M. F. Kanevski, and W. J. Emery, "Active learning methods for remote sensing image classification," IEEE Trans. Geosci. Remote Sens., vol. 47, no. 7, pp. 2218–2232, Jul. 2009.

[13] Y. Gu, Q. Wang, X. Jia, and J. A. Benediktsson, "A novel MKL model of integrating LiDAR data and MSI for urban area classification," IEEE Trans. Geosci. Remote Sens., vol. 53, no. 10, pp. 5312–5326, Oct. 2015.

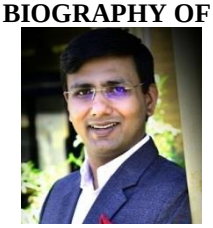
[14] F. Melgani and L. Bruzzone, "Classification of hyperspectral remote sensing images with support vector machines," IEEE Trans. Geosci. Remote Sens., vol. 42, no. 8, pp. 1778–1790, Aug. 2004.

[15] S Chandeesh Kumar, M Hemalatha, S Badri Narayan, P Nandhini, “Region Driven Remote Sensing Image Captioning”, Procedia Computer Science, Elsevier, Volume 165, 2019, Pages 32-40, ISSN 1877-0509, https://doi.org/10.1016/j.procs.2020.01.067.
[16] Z. Shi and Z. Zou, “Can a machine generate humanlike language descriptions for a remote sensing image?” IEEE Transactions Geoscience Remote Sensing, vol. 55, no. 6, pp. 3623–3634, Jun. 2017.
[17] X. Zhang, Q. Wang, S. Chen and X. Li, "Multi-Scale Cropping Mechanism for Remote Sensing Image Captioning," IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium, Yokohama, Japan, 2019, pp. 10039-10042, doi: 10.1109/IGARSS.2019.8900503.
[18] B. Wang, X. Zheng, B. Qu and X. Lu, "Retrieval Topic Recurrent Memory Network for Remote Sensing Image Captioning," in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 13, pp. 256-270, 2020, doi: 10.1109/JSTARS.2019.2959208.
[19] B. Z. Yao, X. Yang, L. Lin, M. W. Lee and S. Zhu, "I2T: Image Parsing to Text Description," in Proceedings of the IEEE, vol. 98, no. 8, pp. 1485-1508, Aug. 2010, doi: 10.1109/JPROC.2010.2050411.
[20] V. Ordonez, G. Kulkarni, and T. L. Berg, “Im2text: Describing images using 1 million captioned photographs,” in Proc. Adv. Neural Inf. Process. Syst., 2011, pp. 1143–1151.
[21] S. Li, G. Kulkarni, T. L. Berg, A. C. Berg, and Y. Choi, “Composing simple image descriptions using Web-scale n-grams,” in Proc. 15th Conf. Comput. Natural Lang. Learn., 2011, pp. 220–228.
[22] Y. Yang, C. L. Teo, D. H. Iii, and Y. Aloimonos, “Corpus-guided sentence generation of natural images,” in Proc. Conf. Empirical Methods Natural Lang. Process., 2011, pp. 444–454.
[23] Y. Feng and M. Lapata, "Automatic Caption Generation for News Images," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 35, no. 4, pp. 797-812, April 2013, doi: 10.1109/TPAMI.2012.118.
[24] G. Kulkarni et al., "BabyTalk: Understanding and Generating Simple Image Descriptions," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 35, no. 12, pp. 2891-2903, Dec. 2013, doi: 10.1109/TPAMI.2012.162.
[25] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. (2014). “Show and tell: A neural image caption generator.” [Online]. Available: https://arxiv.org/abs/1411.4555
[26] A. Karpathy and L. Fei-Fei, “Deep visual-semantic alignments for generating image descriptions,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2015, pp. 3128–3137.
[27] K. Xu et al., “Show, attend and tell: Neural image caption generation with visual attention,” in Proc. Int. Conf. Mach. Learn., 2015, pp. 2048–2057.
[28] J. Johnson, A. Karpathy, and L. Fei-Fei, “DenseCap: Fully convolutional localization networks for dense captioning,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2016, pp. 4565–4574.
[29] O. Vinyals, A. Toshev, S. Bengio and D. Erhan, "Show and Tell: Lessons Learned from the 2015 MSCOCO Image Captioning Challenge," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 4, pp. 652-663, 1 April 2017, doi: 10.1109/TPAMI.2016.2587640.
[30] A. Karpathy and L. Fei-Fei, "Deep Visual-Semantic Alignments for Generating Image Descriptions," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 4, pp. 664-676, 1 April 2017, doi: 10.1109/TPAMI.2016.2598339.
[31] A. Tariq and H. Foroosh, "A Context-Driven Extractive Framework for Generating Realistic Image Descriptions," in IEEE Transactions on Image Processing, vol. 26, no. 2, pp. 619-632, Feb. 2017, doi: 10.1109/TIP.2016.2628585.
[32] K. Fu, J. Jin, R. Cui, F. Sha and C. Zhang, "Aligning Where to See and What to Tell: Image Captioning with Region-Based Attention and Scene-Specific Contexts," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 12, pp. 2321-2334, 1 Dec. 2017, doi: 10.1109/TPAMI.2016.2642953.
[33] C. C. Park, B. Kim and G. KIM, "Towards Personalized Image Captioning via Multimodal Memory Networks," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 41, no. 4, pp. 999-1012, 1 April 2019, doi: 10.1109/TPAMI.2018.2824816.
[34] M. Yang et al., "Multitask Learning for Cross-Domain Image Captioning," in IEEE Transactions on Multimedia, vol. 21, no. 4, pp. 1047-1061, April 2019, doi: 10.1109/TMM.2018.2869276.
[35] N. Yu, X. Hu, B. Song, J. Yang and J. Zhang, "Topic-Oriented Image Captioning Based on Order-Embedding," in IEEE Transactions on Image Processing, vol. 28, no. 6, pp. 2743-2754, June 2019, doi: 10.1109/TIP.2018.2889922.
[36] X. Li and S. Jiang, "Know More Say Less: Image Captioning Based on Scene Graphs," in IEEE Transactions on Multimedia, vol. 21, no. 8, pp. 2117-2130, Aug. 2019, doi: 10.1109/TMM.2019.2896516.
[37] Z. Zha, D. Liu, H. Zhang, Y. Zhang and F. Wu, "Context-Aware Visual Policy Network for Fine-Grained Image Captioning," in IEEE Transactions on Pattern Analysis and Machine Intelligence, doi: 10.1109/TPAMI.2019.2909864.
[38] L. Gao, X. Li, J. Song and H. T. Shen, "Hierarchical LSTMs with Adaptive Attention for Visual Captioning," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 42, no. 5, pp. 1112-1131, 1 May 2020, doi: 10.1109/TPAMI.2019.2894139.
[39] N. Xu et al., "Multi-Level Policy and Reward-Based Deep Reinforcement Learning Framework for Image Captioning," in IEEE Transactions on Multimedia, vol. 22, no. 5, pp. 1372-1383, May 2020, doi: 10.1109/TMM.2019.2941820.
[40] B. Wang, C. Wang, Q. Zhang, Y. Su, Y. Wang and Y. Xu, "Cross-Lingual Image Caption Generation Based on Visual Attention Model," in IEEE Access, vol. 8, pp. 104543-104554, 2020, doi: 10.1109/ACCESS.2020.2999568.
[41] Y. Wang, N. Xu, A. -A. Liu, W. Li and Y. Zhang, "High-Order Interaction Learning for Image Captioning," in IEEE Transactions on Circuits and Systems for Video Technology, doi: 10.1109/TCSVT.2021.3121062.

[42] D. -J. Kim, T. -H. Oh, J. Choi and I. S. Kweon, "Dense Relational Image Captioning via Multi-task Triple-Stream Networks," in IEEE Transactions on Pattern Analysis and Machine Intelligence, doi: 10.1109/TPAMI.2021.3119754.
[43] A. -A. Liu, Y. Zhai, N. Xu, W. Nie, W. Li and Y. Zhang, "Region-Aware Image Captioning via Interaction Learning," in IEEE Transactions on Circuits and Systems for Video Technology, doi: 10.1109/TCSVT.2021.3107035.
[44] C. Yan et al., "Task-Adaptive Attention for Image Captioning," in IEEE Transactions on Circuits and Systems for Video Technology, doi: 10.1109/TCSVT.2021.3067449.
[45] H. Ben et al., "Unpaired Image Captioning with Semantic-Constrained Self-Learning," in IEEE Transactions on Multimedia, doi: 10.1109/TMM.2021.3060948.
[46] S. Zhang, Y. Zhang, Z. Chen and Z. Li, "VSAM-Based Visual Keyword Generation for Image Caption," in IEEE Access, vol. 9, pp. 27638-27649, 2021, doi: 10.1109/ACCESS.2021.3058425.
[47] Z. Zhou et al., "An Image Captioning Model Based on Bidirectional Depth Residuals and its Application," in IEEE Access, vol. 9, pp. 25360-25370, 2021, doi: 10.1109/ACCESS.2021.3057091.
[48] L. Huo, L. Bai and S. -M. Zhou, "Automatically Generating Natural Language Descriptions of Images by a Deep Hierarchical Framework," in IEEE Transactions on Cybernetics, doi: 10.1109/TCYB.2020.3041595.
[49] X. Lu, B. Wang, X. Zheng and X. Li, "Exploring Models and Data for Remote Sensing Image Caption Generation," in IEEE Transactions on Geoscience and Remote Sensing, vol. 56, no. 4, pp. 2183-2195, April 2018.
[50] S. Wu, X. Zhang, X. Wang, C. Li and L. Jiao, "Scene Attention Mechanism for Remote Sensing Image Caption Generation," 2020 International Joint Conference on Neural Networks (IJCNN), Glasgow, UK, 2020, pp. 1-7, doi: 10.1109/IJCNN48605.2020.9207381.
[51] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv preprint arXiv:1409.1556, 2014.
[52] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman. (2014). "Return of the devil in the details: Delving deep into convolutional nets." [Online]. Available: https://arxiv.org/abs/1405.3531
[53] B. Wang, X. Lu, X. Zheng and X. Li, "Semantic Descriptions of High-Resolution Remote Sensing Images," in IEEE Geoscience and Remote Sensing Letters, vol. 16, no. 8, pp. 1274-1278, Aug. 2019, doi: 10.1109/LGRS.2019.2893772.
[54] B. Qu, X. Li, D. Tao, and X. Lu, "Deep semantic understanding of high resolution remote sensing image," in Proc. Int. Conf. Comput., Inf. Telecommun. Syst., Jul. 2016, pp. 124–128.
[55] X. Zhang, X. Li, J. An, L. Gao, B. Hou and C. Li, "Natural language description of remote sensing images based on deep learning," 2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), 2017, pp. 4798-4801, doi: 10.1109/IGARSS.2017.8128075.
[56] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in Advances in neural information processing systems, 2012, pp. 1097–1105.
[57] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1–9.
[58] Y. Jia. Caffe: An Open Source Convolutional Architecture for Fast Feature Embedding. http://caffe.berkeleyvision.org/,2013.
[59] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 37, no. 9, pp. 1904–1916, 2015.
[60] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.
[61] S. Wang, J. Chen, and G. Wang, "Intensive positioning network for remote sensing image captioning," in Proc. Int. Conf. Intell. Sci. Big Data Eng., 2018, pp. 567–576.
[62] X. Zhang, X. Wang, X. Tang, H. Zhou, and C. Li, "Description Generation for Remote Sensing Images Using Attribute Attention Mechanism," Remote Sensing, vol. 11, no. 6, p. 612, Mar. 2019 [Online]. Available: http://dx.doi.org/10.3390/rs11060612
[63] X. Lu, B. Wang and X. Zheng, "Sound Active Attention Framework for Remote Sensing Image Captioning," in IEEE Transactions on Geoscience and Remote Sensing, vol. 58, no. 3, pp. 1985-2000, March 2020, doi: 10.1109/TGRS.2019.2951636.
[64] X. Shen, B. Liu, Y. Zhou, J. Zhao, and M. Liu, "Remote sensing image captioning via variational autoencoder and reinforcement learning," Knowl. Based Syst., vol. 203, 2020, Art. no. 105920.
[65] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," Int. J. Comput. Vis., vol. 60, no. 2, pp. 91–110, 2004.
[66] J. Sivic and A. Zisserman, "Video Google: A text retrieval approach to object matching in videos," in Proc. 9th IEEE Int. Conf. Comput. Vis., vol. 2. Oct. 2003, pp. 1470–1477.
[67] F. Perronnin, Y. Liu, J. Sánchez, and H. Poirier, "Large-scale image retrieval with compressed fisher vectors," in Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., Jun. 2010, pp. 3384–3391.
[68] H. Jégou, F. Perronnin, M. Douze, J. Sánchez, P. Pérez, and C. Schmid, "Aggregating local image descriptors into compact codes," IEEE Trans. Pattern Anal. Mach. Intell., vol. 34, no. 9, pp. 1704–1716, Sep. 2012.
[69] Tomas Mikolov and Kai Chen and Greg S. Corrado and Jeffrey Dean (2013) "Efficient Estimation of Word Representations in Vector Space", arXiv:1301.3781
[70] G. Cheng, J. Han, X. Lu, Remote sensing image scene classification: Benchmark and state of the art, Proc. IEEE 105 (10) (2017) 1865–1883, http://dx.doi.org/10.1109/jproc.2017.2675998.

[71] G. Hoxha, F. Melgani and J. Slaghenauffi, "A New CNN-RNN Framework For Remote Sensing Image Captioning," 2020 Mediterranean and Middle-East Geoscience and Remote Sensing Symposium (M2GARSS), Tunis, Tunisia, 2020, pp. 1-4, doi: 10.1109/M2GARSS47143.2020.9105191.
[72] J. Devlin, S. Gupta, R. Girshick, M. Mitchell, and C. L. Zitnick, "Exploring Nearest Neighbor Approaches for Image Captioning", ArXiv 150504467 Cs, May 2015.
[73] M. Hodosh, P. Young, and J. Hockenmaier, "Framing image description as a ranking task: Data, models and evaluation metrics," J. Artif. Intell. Res., vol. 47, no. 8, pp. 853–899, 2013.
[74] A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, and T. Mikolov, "Devise: A deep visual-semantic embedding model," in Proc. Adv. Neural Inf. Process. Syst., 2013, pp. 2121–2129.
[75] Yi Yang and Shawn Newsam, "Bag-Of-Visual-Words and Spatial Extensions for Land-Use Classification," ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (ACM GIS), 2010.
[76] F. Zhang, B. Du, and L. Zhang, "Saliency-guided unsupervised feature learning for scene classification," Geoscience and Remote Sensing, IEEE Transactions on, vol. 53, no. 4, pp. 2175–2184, 2015.
[77] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier, "From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions," Trans. Assoc. Comput. Linguistics, vol. 2, pp. 67–78, 2014.
[78] X. Chen et al. (2015). "Microsoft COCO captions: Data collection and evaluation server." [Online]. Available: https://arxiv.org/abs/1504.00325
[79] G. S. Xia et al., "AID: A benchmark data set for performance evaluation of aerial scene classification," IEEE Trans. Geosci. Remote Sens., vol. 55, no. 7, pp. 3965–3981, Jul. 2016.
[80] Dhaksha Team (2018) "Drone Manufacture in INDIA - Team Dhaksha", https://www.teamdhaksha.com/.
[81] K. Papineni, S. Roukos, T. Ward, and W. J. Zhu, "BLEU: A method for automatic evaluation of machine translation," in Proc. 40th Annu. Meet. Assoc. Comput. Linguistics, 2002, pp. 311–318.
[82] C. Flick, "Rouge: A package for automatic evaluation of summaries," in Proc. Workshop Text Summarization Branches Out, 2004, p. 10.
[83] M. Denkowski and A. Lavie, "Meteor universal: Language specific translation evaluation for any target language," in Proc. Workshop Stat. Mach. Transl., 2014, pp. 376–380.
[84] R. Vedantam, C. L. Zitnick, and D. Parikh, "CIDEr: Consensus-based image description evaluation," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2015, pp. 4566–4575.
[85] P. Anderson, B. Fernando, M. Johnson, and S. Gould, "Spice: Semantic propositional image caption evaluation," in Proc. Eur. Conf. Comput. Vis., 2016, pp. 382–398.

BIOGRAPHY OF AUTHORS



Parag J. Mondhe is Ph.D. Research Scholar at Matoshri College of Engineering and Research Centre, Nashik, India. He is working as an Assistant Professor at K. K. Wagh Institute of Engineering Education and Research, Nashik, India since 2014. He has completed Master of Engineering and Bachelor of Engineering from Savitribai Phule Pune University, India in 2014 and 2012 respectively. His research papers are published in proceedings of international conferences and journals indexed by Scopus. His area of interest includes signal processing, artificial intelligence and embedded systems.

Dr. Manisha P. Satone is a Professor of Electronics and Telecommunication Engineering at Matoshri College of Engineering and Research Centre, Nashik, India. She was awarded Ph.D. by Savitribai Phule Pune University, India in 2015. She has teaching experience of more than 31 years. Her research papers are published in journals indexed by Web of Science, Scopus. She has fetched a research grant and acquired copyrights and patents. She has provided consultancy to many multinational companies. Her area of expertise includes signal processing, artificial intelligence and embedded systems.

Dr. Gajanan K. Kharate is a Principal at Matoshri College of Engineering and Research Centre, Nashik, India. He is a former Dean, Faculty of Engineering, Savitribai Phule Pune University, India. He was awarded Ph.D. by Savitribai Phule Pune University, India in 2007. He has teaching experience of more than 36 years. His research papers are published in journals indexed by Web of Science, Scopus. He has fetched a research grant and acquired copyrights and patents. His book is published by Oxford University Press. His area of expertise includes signal processing, communication engineering, VLSI technology.