An Encoder-Decoder Framework for Remote Sensing Image Captioning

MSc Research Project
Data Analytics

Namita Mohan
Student ID: 19212500

School of Computing
National College of Ireland

Supervisor: Dr Paul Stynes and Dr Pramod Pathak

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Namita Mohan
Student ID:	19212500
Programme:	Data Analytics
Year:	2020/2021
Module:	MSc Research Project
Supervisor:	Dr Paul Stynes and Dr Pramod Pathak
Submission Due Date:	16/08/2021
Project Title:	An Encoder-Decoder Framework for Remote Sensing Image Captioning
Word Count:	4260
Page Count:	12

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Namita Mohan
Date:	19th September 2021

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

An Encoder-Decoder Framework for Remote Sensing Image Captioning

Namita Mohan
19212500

Abstract

Remote sensing image captioning involves generating human-like text to describe the content of images representing the earth's surface captured from a distance. The challenge with remote sensing image captioning is identifying multiple objects with large scale differences and the relationship between these objects present in a single image. This research proposes a novel encoder-decoder framework to identify objects and their relationship for generating human-like descriptions of remote sensing images. The encoder-decoder framework consists of Capsule Network and Bidirectional LSTM. The Capsule Network is used to extract the object features from the images and Bi-directional LSTM is used to generate text descriptions based on the object features. This research used the popular RSICD dataset with over ten thousand remote sensing images for the training of the proposed framework. The implemented frameworks improved the BLEU-1 by 7%, BLEU-2 by 16%, BLEU-3 by 20% and BLEU-4 by 17% as compared to the traditional CNN-RNN model. The results demonstrate the power of Capsule Network compared to CNN however, the performance of the framework can be improved by using deeper architecture Capsule Network and much larger datasets. Remote sensing image captioning would be useful for image retrieval in all applications of remote sensing images such as development planning and disaster monitoring.

1 Introduction

The advancement of technology and image acquiring devices led to increase in the availability of high spatial resolution remote sensing images. After image classification, object recognition, scene classification and image segmentation many researchers are showing interest in image captioning as it provides more information about the content of an image (Lu et al.; 2018; Shi and Zou; 2017; Qu et al.; 2016). As remote sensing image captioning is a multimodality problem, it involves both natural language processing and computer vision. Therefore, it is a challenging problem to solve. Automatically generated text descriptions for remote sensing images can significantly improve image retrieval systems by including more words other than keywords for searching images. Study of the earth surface, disaster monitoring, development planning and military intelligence generation are some fields where remote sensing image captioning is useful (Shi and Zou; 2017).

There are three types of image captioning techniques: retrieval-based, object detection-based, and encoder-decoder network based (Yuan et al.; 2020). Image captioning methods based on retrieval approach look for images similar to the query image and then generate sentences based on the description of similar images. However, sentences formed using retrieval-based methods have a limited semantic range and may be irrelevant if no same type of image is identified in the archive. The object detction approach uses three types of information from images: objects, visual properties of objects and relationships between these objects to generate a text describing the target image. Object detection-based descriptions are comparatively simple and can be improper and incorrect, which needs to be avoided. Encoder-decoder network-based captioning approaches employ a mix of neural networks to extract visual elements from an image depending on its content, and then use this knowledge to construct a coherent and semantically correct description. It is observed that an encoder-decoder model outperforms retrieval-based and object detection-based approaches for image captioning problem (Yuan et al.; 2020). A convolutional neural network (CNN) and a recurrent neural network (RNN) or long short-term memory (LSTM) network are typically employed to generate text descriptions based on image features (Qu et al.; 2016).

The main drawback of using CNN in utilizing it to extract features from remote sensing images is the inability to handle rotational invariance. Therefore, to achieve better performance it must be explicitly trained on various augmentations of the single image like crop, zoom, invert and different transformations. The use of CNN requires not only a large dataset for training but also augmentation which is a costly computationally task. In addition, due to the hierarchy of the CNN parts, the information about the relationship between objects present in the image is lost while forwarding information to detect different objects in an image. To overcome the drawbacks of CNN, this study proposes the usage of a Capsule Network, because of its ability to capture both position and orientation of an entity present in an image with the help of activation vectors. Instead of MaxPooling in CNN, Capsule Network uses Dynamic Routing between its capsules which makes images with viewpoint invariance understandable. In addition, the hierarchy of parts in Capsule Network is opposite to CNN, which means that the current layer tries to predict the probability of detecting an object in the next layer. Therefore, the output of the current layer goes to that capsule in the next layer in which the probability of finding that object is maximum. RNN is a popular choice for sequence generation problems however, because of the vanishing gradient problem it faces limitations in generating long sentences to describe an image (Wang et al.; 2020). Therefore, unidirectional LSTM with the power of the memory cell is used to get better results in image captioning problems. This research proposes the use of bi-directional LSTM to improve sentence generation as it takes advantage of both past and future information while predicting.

This study tries to answer the research question, to what extent can an encoder-decoder framework be used for generating human-like text to describe remote sensing images adopting Capsule Network and Bi-directional LSTM. For implementation, Remote Sensing Image Captioning Dataset (RSICD) is used, which contains images and five reference sentences for each image. RSICD dataset is specially designed for remote sensing image captioning with the help of professionals from the domain. The objectives of this research are to investigate the state of the art researches in the remote sensing image captioning domain, design and implement a novel encoder-decoder framework using capsule network and bi-directional LSTM and critical evaluation of implemented frameworks using BLEU and ROUGE metrics.

The major contribution of this study is a novel encoder-decoder framework for generating human-like text targeted to better describe entities and their relation to each other, present in high spatial resolution remote sensing images as compare to the state of the art solutions.

The rest of this research paper is as follows: Section 2 discusses the work done related to remote sensing image captioning, natural image captioning, capsule network as feature extractor and bi-directional LSTM for text sequence generation. The next Section 3 illustrate the adopted methodology for this research. The following Section 4 presents the design specifications of the implemented models. Section 5 provides details about the implementation. Section 6 presents the evaluation of developed frameworks on the basis of selected evaluation metrics and discuss findings. Finally, The conclusion of this research and future work is provided in Section 7.

2 Related Work

The process of automatically generating human-like text to describe the content of a remote sensing image in a comprehensible context is known as remote sensing image captioning. In comparison to retrieval-based and object detection-based approaches, the encoder-decoder based method has been demonstrated to produce better and correct sentences with varying length and syntax (Zhang, Diao, Zhang, Yan, Gao and Sun; 2019). This section discusses researches conducting the domain of remote sensing image captioning to identify best practices for guiding decisions of this research and to explain the limitations of current solutions.

2.1 Natural Image Captioning

Natural image captioning intent to generate human-like text for describing the content of natural images that presents the front view of objects. The inspiration for image captioning is from how humans recognize their surroundings in a single glance. Recently many researchers explored image captioning using deep learning models such as CNN, RNN and LSTM. The study Xu et al. (2020) uses a novel framework based on reinforcement learning (RL) to solve the image captioning problem. They showed as image captioning involves multiple modalities (vision and language) it requires a multi-level reward policy. In another study by Zhou et al. (2020) a novel multi-level visual fusion network based on RL is used to improve the relationship between entities and attributes of an image in the generated descriptions. Which includes three parts: visual network to extract visual features, fusion network to focus on non-visual information and language network of LSTM to generate sentences. Akilan et al. (2020) studies how a basic image captioning model reacts to visual and textual clues to understand the multimodal representations in captioning models. They showed significant improvement in the model by comparing unimodal frameworks to multimodal frameworks with various fusion techniques such as tensor-based, attention-based and operation-based.

2.2 Remote Sensing Image Captioning

The traditional remote sensing image captioning frameworks have two major limitations of not generating new descriptions and misrecognition of objects and scenes in the image. Therefore, to resolve these two problems Hoxha, Melgani and Slaghenauffi (2020) proposed a novel framework by combining the retrieval-based method with the traditional CNN-RNN encoder-decoder framework with an aim to improve the sentences describing remote sensing images. They used the RSICD dataset for training and testing the framework. In another study Hoxha, Melgani and Demir (2020), a novel system is proposed to retrieve similar images from the database with the help of CNN-RNN based image captioning framework. Huang et al. (2021) used a denoising-based multi-scale feature fusion (DMSFF) mechanism that extracted image features from a different layer of CNN based encoder model which then passed through denoising channels before merging to obtain the final feature vector. A Variational Autoencoder and Reinforcement Learning based Two-stage Multi-task Learning Model (VRTMM) proposed by Shen et al. (2020) resolve the problem of overfitting and improved semantic information in generated description. Based on various evaluation metrics, this model performed better than the simple encoder-decoder framework for captioning. Inspired by cropping mechanism used in image classification Zhang, Wang, Chen and Li (2019) enhanced the performance of CNN encoder by multi-scale cropping of the input image. They used a combination of three different pre-trained CNN models VGG16, Inception-ResNetV2 and ResNet-152 for extracting image features and LSTM network for generating sentences.

2.3 Attention based methods

Attention-based methods tries to replicate the way humans analyze visual information in just a single glance. They aim to extract more information about relationship between the entities presented in the image to include human perspective in the generate sentences. In the study Wu et al. (2020), an attention vector is used with the traditional CNN-LSTM encoder-decoder framework which achieved better performance than other popular models. Inspired from convolutional attention, a novel Label-Attention Mechanism (LAM) proposed in Zhang, Diao, Zhang, Yan, Gao and Sun (2019) which uses label information in the calculation of attention masks to focus on different parts based on the content of the input image. In another study Zhang, Wang, Tang, Zhou and Li (2019), a fully connected CNN layer based on VGG16 is used as an attention layer to focus on specific parts of the image. The attention methods for the natural image were not performing well with the complexity of remote sensing images therefore, in a research Li et al. (2020) a multi-level attention mechanism used which successfully identified the attention type in the remote sensing image. Yuan et al. (2020) used multi-level attention to resolve the problem of variable scale in the remote sensing images. Along with CNN encoder and multi-level feature maps, graph convolutional networks and pre-computed attribute graphs are used to generate attribute features.

2.4 Capsule Network and Bidirectional LSTM

Capsule Network overcomes the drawbacks of CNN such as rotational invariance and information loss because of max-pooling by using a group of neurons called a capsule and dynamic routing between these capsules (Sabour et al.; 2017). Because of its ability to capture the existence and orientation of objects present in an image, the capsule network has proven to provide significantly improved results in object detection problems. Guo et al. (2021) showed how Capsule Network in U-net model performed better than traditional CNN model in the classification of remote sensing images. Deka (2020) compared the performance of Capsule Network and CNN as an encoder for feature extraction in an encoder-decoder framework for natural image captioning. The study showed the combination of VGG16 and Capsule Network provided the best BLEU (n=1,2,3,4) scores on the benchmark dataset Flickr8k.

Long Short Term Memory Network is a popular choice for generating text sequences. With the power of memory cells, LSTM has proven to provide better results than RNN in image captioning problems. Inspired from LSTM, Bi-directional Long Short Term Memory Network (BiLSTM) introduced in Schuster and Paliwal (1997) which contain two hidden layers to use past and future information while predicting current words. Siamese difference captioning model (SDCM) introduced in Oluwasanmi et al. (2019) successfully used Bi-directional LSTM to generate text describing the difference between two images. In the study Agughalam et al. (2021), a Bi-directional LSTM based decoder is used to generate human-like text for describing natural images with InceptionV3 for extracting global features and VGG16Places365 for extracting scene information. The research concludes that BiLSTM with scene features improved the quality and correctness of the generated descriptions, compared to the LSTM based encoder-decoder image captioning framework.

In conclusion, it is proven that an encoder-decoder framework using CNN-RNN or CNN-LSTM produce meaningful sentences compared to the other captioning methods. However, due to CNN's limitations, the descriptions lack to represent the relationship between the objects in the image. Also, uni-directional LSTM has a limitation of using only past information for predicting the current word. Aiming to overcome these limitations of CNN and LSTM this study investigates an encoder-decoder framework by leveraging the power of Capsule Network and BiLSTM to generate semantically correct and meaningful human-like text descriptions for remote sensing images. Moreover, the combination of Capsule Network and BiLSTM does not appear to be used in any previous work for remote sensing image captioning.

3 Methodology

The research methodology includes data collection, data processing, data transformation, modelling, evaluation and results as illustrated in Figure 1. In the first step Data Collection, the Remote Sensing Image Captioning Dataset (RSICD) containing in total 10,921 images each of which have five reference sentences is selected. There are a total of 3323 distinct words and 24,333 sentences in the dataset (Lu et al.; 2018). To get five sentences for each image, some captions are repeated randomly from the existing captions. Figure 2 represents two sample images along with captions from RSICD.

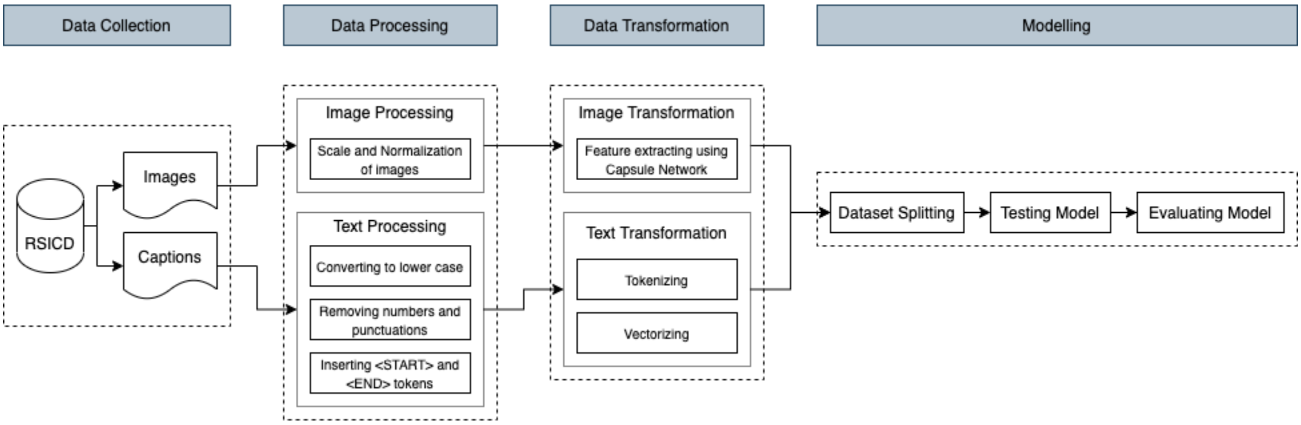
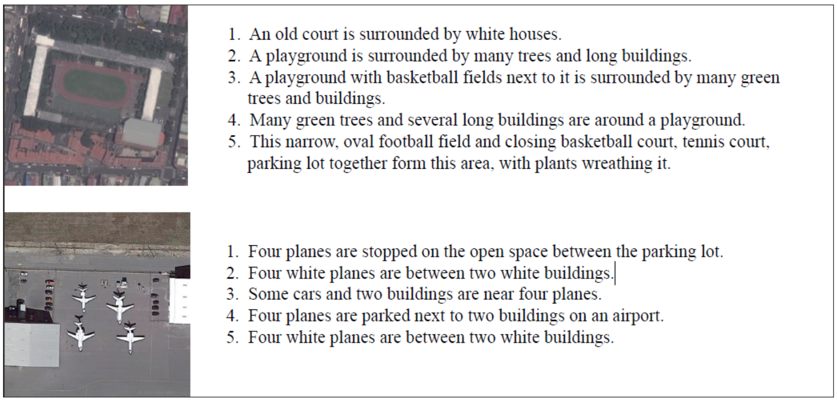


Figure 1: Research Methodology.

In the second step Data Processing, image and text processing are performed to data get data ready for modelling. Image processing involves loading images and resizing them to the input shape of InceptionV3 and Capsule Network encoder models for image feature extraction. In text processing, first text descriptions are converted to lower cases as words with different capitalizations can increase vocabulary size unnecessary. Second, punctuations and numerical data are removed from the descriptions. At last, '<START>' and '<END>' tokens are inserted into each description to indicate starting and end of the description respectively.



1. An old court is surrounded by white houses.
2. A playground is surrounded by many trees and long buildings.
3. A playground with basketball fields next to it is surrounded by many green trees and buildings.
4. Many green trees and several long buildings are around a playground.
5. This narrow, oval football field and closing basketball court, tennis court, parking lot together form this area, with plants wreathing it.

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 2: Sample images and reference sentences from RSICD dataset.

In Data Transformation, image and text data is transformed into a form required for modelling. Image transformation includes extracting features from images and converts them to vectors for further processing. Inception v3 CNN model pre-trained on ImageNet dataset and the purposed Capsule network are used as encoders for extracting image features. The output of the second last layer of these models is used as a feature vector for the decoder. Text transformation includes tokenization followed by vectorization of the image descriptions. First, each word in the text descriptions is replaced by a numeral value based on the dictionary to get word to index and inverse dictionary to get the index to word. Then, the maximum length text vector is looked out to get the text vectors of equal size. All shorter text sequences are then padded with zeros to get the same length text vectors.

In the Modelling step, the RSICD dataset is split into three parts of 8734, 1094 and 1093 images for training, validation and testing respectively. The implemented encoder-decoder framework requires two inputs, the feature vector of the image and word vectors of the actual text descriptions and output text sequence. First, feature vectors of images are generated using the encoder model which is passed along with tokenized word vector to the decoder model. The decoder model concatenates the image feature vector model and word vector, then passed to fully connected LSTM or Bi-LSTM layers. In the end, the decoder model has two dense layers first with relu activation function and second with softmax activation which predicts the probability of each word next in the sequence. The sentences are generated using the greedy approach for inference. That means the word with the highest probability is selected as the next word in the sequence. For the evaluation of the implemented framework two metrics BLEU and ROUGE-L are selected as they are most common in the researches discussed in section 2.

4 Design Specification

Figure 3 illustrate the model architecture including input layers, embedding layers, dropout layers, dense layers and bidirectional LSTM layer. The first input layer takes image feature vectors with input shape depends on the length of the output vector from the encoder model. For the InceptionV3 CNN based encoder model, the input shape was 2048 and for the Capsule Network based encoder model the input shape was 320. A dropout layer is added to reduce overfitting after the first input layer, followed by a dense layer to reduce the size of image feature input equal to the size of text input that is 256. The second input layer is used for taking word vector therefore, the input shape depends on the maximum length of the word vector which is 36 for the implemented framework. The text input layer is followed by an embedding layer to learn vector respective is each word appears in the text descriptions. The output of the embedding layer is a vector of length 256 which is then passed to a dropout layer.

The outputs from input layers for the image feature and for the word vector are concatenated together and then passed to a bidirectional LSTM layer of size 512 units. Which learns the occurrence current word based on the last predicted and next word in the sequence. The BiLSTM layer is followed by two dense layers first with activation function relu, input shape 512 and output shape of 256. The second dense layer has input shape 256, output shape 3093 and softmax activation function. The final softmax layer predicts the probability of each word to the next in the sequence.

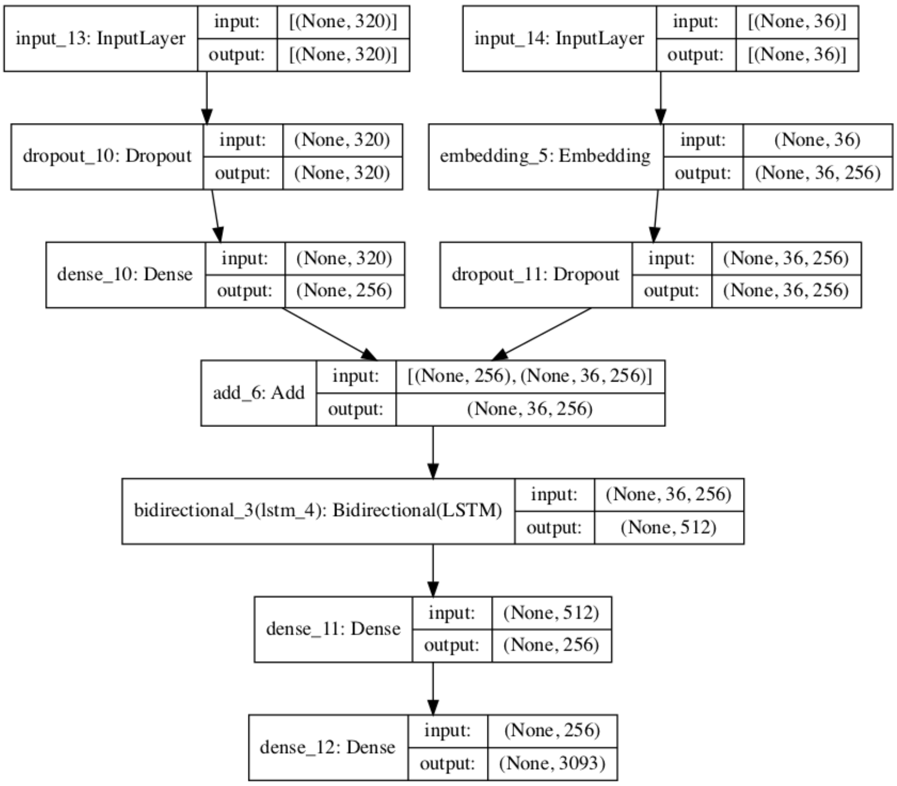


Figure 3: Model Architecture.

5 Implementation

For the implementation of this research, a MacBook Pro with an Intel Core i5 processor and 8GB RAM is used. Jupyter notebook is used to execute all codes in Python 3. The dataset consists of images in JPG format and a JSON file with the image name as key and array reference descriptions as value. Both images and text descriptions are loaded and preprocessed in the notebook as described in the section 3. For training, the model 8734 images are used and 1094 images are used for the validation. 'categorical_crossentropy' and 'RMSProp' are used as the loss and backpropagation to optimize the training as this proven to provide efficient and fast convergence in problems of neural language translations Agughalam et al. (2021). Early stopping and checkpoints are used to avoid overfitting and save the model only when the validation loss improves. A total of 10 epochs are used for the training of each model and BLEU (n=1,2,3,4) and ROUGE-L evaluation metrics are used for monitoring overfitting.

6 Evaluation

The outcomes of the experiments conducted in this research will be discussed in this section. This research includes in total three experiments. The first experiment is an implementation of the state of the art solution for remote sensing image captioning, the second experiment involves employing Capsule Network as an encoder for feature extraction along with BiLSTM for text generation and the final experiment involves merging both CNN and Capsule Network as an encoder for feature extraction. For inference, this research used greedy search as the most applied in pervious researches and computationally fast. The results from these experiments are compared based on BLEU and ROUGE evaluation metrics. BLEU is a traditional metric that uses geometric mean to compare each word predicted sentence to the reference sentences. Like BLEU, ROUGE also matches n-grams for comparing sentences. However, it is based on recall, while BLEU is based on precision.

6.1 Experiment 1

This experiment implements the state of the art CNN-LSTM encoder-decoder framework in remote sensing image captioning Hoxha, Melgani and Slaghenauffi (2020); Wu et al. (2020). The image features are extracted using the inceptionV3 model, a CNN model pre-trained on the ImageNet dataset. For generating text sequences the LSTM model is used in the decoder part. The model was trained for 10 epochs with an improvement in validation loss. As presented in Table 1, the BLUE-1 score is 87.45, the BLUE-2 score is 76.64, the BLUE-3 score is 69.10, the BLUE-4 score is 58.68 and the ROUGE-L score is 46.01. The results from this experiment showed significant results as compared to the previous researches where the CNN model is trained from scratch on remote sensing images. The next experiment is conducted to observe the performance of Capsule Network as compared to CNN.

6.2 Experiment 2

This experiment extends the state of the art model by using Capsule Network instead of CNN for extracting features from the images. The architecture of implemented Capsule

Network is inspired by the study (Deka; 2020). The input layer of the encoder model is a convolutional 2D layer of size 128 which then connects to a capsule layer consist 10 capsules of dimension 32. The output of these capsules are concatenated together and passed to flatten layer which generates an image feature vector of length 320. For the decoder part, a BiLSTM is used inspired from the research Agughalam et al. (2021). The model was trained for 10 epochs same as the previous experiment. The model achieved an evaluation score of 73.37 in BLEU-1, 64.52 in BLEU-2, 57.63 in BLEU-3, 48.12 in BLEU-4 and 39.06 in ROUGE-L. As Capsule Network is recently introduced and not much explored, simple architecture is followed in this study. The results of this experiment are good enough in comparison to the results of the last experiment.

6.3 Experiment 3

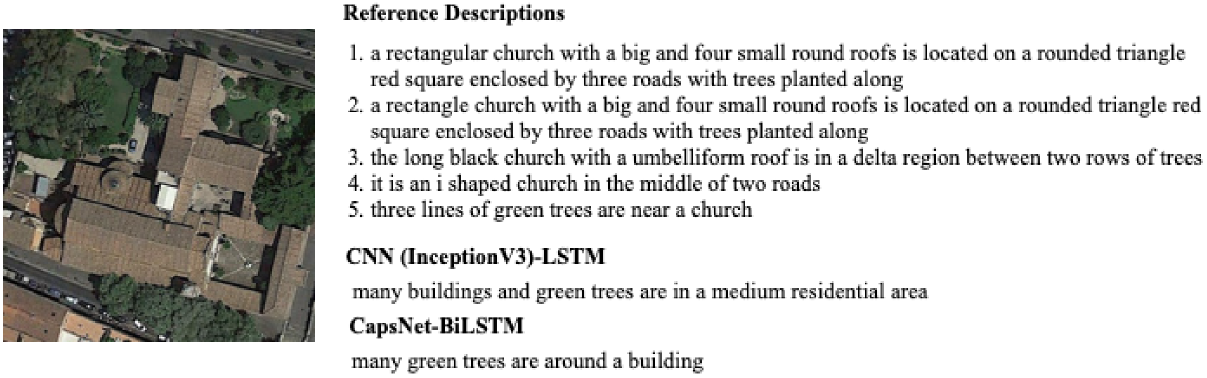
This experiment extends the previous experiment by using features extracting by CNN encoder and CapsNet encoder with BiLSTM decoder. This experiment aimed to observe the result of combining CNN and CapsNet image feature to generate captions. The model was trained for 10 epochs however, the validation loss was not improved after the first epochs which shows the signs of overfitting. As this research used only the best model for generating text descriptions, the trained model after the first epoch is used here. The overfitting also appears in the evaluation scores with BLEU-1 0.019, BLEU-2 0.016, BLEU-3 0.014, BLEU-4 0.012 and ROUGE-L 10.9.

Table 1: BLEU (n=1,2,3,4) and ROUGE-L scores of each experiment

Model	**BLEU-1**	**BLEU-2**	**BLEU-3**	**BLEU-4**	**ROUGE-L**
CNN(InceptionV3)-LSTM	87.45	76.64	69.10	58.68	46.01
CapsNet-BiLSTM	73.37	64.52	57.63	48.12	39.06
CNN-CapsNet-BiLSTM	0.019	0.016	0.014	0.012	10.9

6.4 Discussion

The output of the frameworks implemented in this study presents the power of Capsule Network and BiLSTM to generate meaningful sentences describing the content of remote sensing images. The first experiment follows the architecture of the research done Hoxha, Melgani and Slaghenauffi (2020), however rather than training inceptionV3 CNN on remote sensing images this research used a pre-trained InceptionV3 model for feature extraction. The results from this experiment show the increment of 21% in BLEU-1 and 30% in BLEU-4 as compared to the results of study Hoxha, Melgani and Slaghenauffi (2020). The increase in the evaluation scores explains the power of training models on large datasets. The second experiment as well improves the results by 7% in BLEU-1, 16% in BLEU-2, 20% in BLEU-3 and 17% in BLEU-4 in comparison with the results of the study Hoxha, Melgani and Slaghenauffi (2020). This experiment used a single capsule layer of 10 capsules trained from scratch on the RSICD dataset, illustrating capsule networks' ability in feature extraction of remote sensing images. The third experiment explores the combination of CNN and Capsule Network for feature extraction however, the model does not perform well because of the overfitting on the training dataset. This indicates the requirement of a much large dataset for the training of such a complex model. Figure 4 presents comparison of generated text descriptions by CNN-LSTM and CapsNet-BiLSTM model.



Reference Descriptions

1. a rectangular church with a big and four small round roofs is located on a rounded triangle red square enclosed by three roads with trees planted along
2. a rectangle church with a big and four small round roofs is located on a rounded triangle red square enclosed by three roads with trees planted along
3. the long black church with a umbelliform roof is in a delta region between two rows of trees
4. it is an i shaped church in the middle of two roads
5. three lines of green trees are near a church

CNN (InceptionV3)-LSTM

many buildings and green trees are in a medium residential area

CapsNet-BiLSTM

many green trees are around a building

Figure 4: Example of a figure caption.

7 Conclusion and Future Work

The aim of this research is to investigate to what extent can an encoder-decoder framework be used for generating human-like text to describe remote sensing images using Capsule Network and Bi-directional LSTM. Three different frameworks were implemented in this research for the captioning of remote sensing images. The implemented models were trained on benchmarking dataset RSICD and evaluate on the basis of BLEU and ROUGE-L metrics using the greedy approach for inferencing. This research proposes a novel approach to generate human-like text descriptions for remote sensing images using a combination of CapsNet and BiLSTM in an encoder-decoder framework. The encoder is based on CapsNet consist of 10 capsules which is used for extracting features from the remote sensing images. The key findings of this research is the ability of CapsNet in feature extraction from remote sensing images and the power of large dataset for training to generate better and long sentences. This can be attributed to the capacity of CapsNet in learning image features without image argumentation as required by CNN. Also, the use of BiLSTM shows promising results in improving generated sentences semantically.

The CapsNet used in this research is much smaller than the state of the art CNN model. In future, the deeper architecture of CapsNet can be used with the much larger dataset to improve the performance and reduce the problem of overfitting in encoder-decoder frameworks. Also, various attention mechanisms can be added to this study which can further improve the relationship between objects present in the images. For a complete evaluation of the framework, generated descriptions can be compared on the basis of semantic meaning rather than just n-gram matching.

References

Agughalam, D., Pathak, P. and Stynes, P. (2021). Bidirectional LSTM approach to image captioning with scene features, *in* X. Jiang and H. Fujita (eds), *Thirteenth International Conference on Digital Image Processing (ICDIP 2021)*, Vol. 11878, International

Society for Optics and Photonics, SPIE, pp. 81 – 88. doi: 110.1117/12.2600465.
URL: *https://doi.org/10.1117/12.2600465*

Akilan, T., Thiagarajan, A., Venkatesan, B., Thirumeni, S. and Chandrasekaran, S. G. (2020). Quantifying the impact of complementary visual and textual cues under image captioning, *2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pp. 389–394. doi: 10.1109/SMC42975.2020.9283183.

Deka, J. (2020). *Image captioning: Capsule network vs cnn approach*, Master's thesis, Dublin, National College of Ireland.
URL: *http://norma.ncirl.ie/4298/*

Guo, Y., Liao, J. and Shen, G. (2021). A deep learning model with capsules embedded for high-resolution image classification, *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* **14**: 214–223. doi: 10.1109/JSTARS.2020.3032672.

Hoxha, G., Melgani, F. and Demir, B. (2020). Toward remote sensing image retrieval under a deep image captioning perspective, *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* **13**: 4462–4475. doi: 10.1109/JSTARS.2020.3013818.

Hoxha, G., Melgani, F. and Slaghenauffi, J. (2020). A new cnn-rnn framework for remote sensing image captioning, *2020 Mediterranean and Middle-East Geoscience and Remote Sensing Symposium (M2GARSS)*, pp. 1–4. doi: 10.1109/M2GARSS47143.2020.9105191.

Huang, W., Wang, Q. and Li, X. (2021). Denoising-based multiscale feature fusion for remote sensing image captioning, *IEEE Geoscience and Remote Sensing Letters* **18**(3): 436–440. doi: 10.1109/LGRS.2020.2980933.

Li, Y., Fang, S., Jiao, L., Liu, R. and Shang, R. (2020). A multi-level attention model for remote sensing image captions, *Remote Sensing* **12**(6). doi: 10.3390/rs12060939.

Lu, X., Wang, B., Zheng, X. and Li, X. (2018). Exploring models and data for remote sensing image caption generation, *IEEE Transactions on Geoscience and Remote Sensing* **56**(4): 2183–2195. doi: 10.1109/TGRS.2017.2776321.

Oluwasanmi, A., Aftab, M. U., Alabdulkreem, E., Kumeda, B., Baagyere, E. Y. and Qin, Z. (2019). Captionnet: Automatic end-to-end siamese difference captioning model with attention, *IEEE Access* **7**: 106773–106783. doi: 10.1109/ACCESS.2019.2931223.

Qu, B., Li, X., Tao, D. and Lu, X. (2016). Deep semantic understanding of high resolution remote sensing image, *2016 International Conference on Computer, Information and Telecommunication Systems (CITS)*, pp. 1–5. doi: 10.1109/CITS.2016.7546397.

Sabour, S., Frosst, N. and Hinton, G. E. (2017). Dynamic routing between capsules.

Schuster, M. and Paliwal, K. K. (1997). Bidirectional recurrent neural networks, *IEEE Transactions on Signal Processing* **45**(11): 2673–2681. doi: 10.1109/78.650093.

Shen, X., Liu, B., Zhou, Y., Zhao, J. and Liu, M. (2020). Remote sensing image captioning via variational autoencoder and reinforcement learning, *Knowledge-Based Systems* **203**: 105920. doi: 10.1016/j.knosys.2020.105920.

Shi, Z. and Zou, Z. (2017). Can a machine generate humanlike language descriptions for a remote sensing image?, *IEEE Transactions on Geoscience and Remote Sensing* **55**(6): 3623–3634. doi: 10.1109/TGRS.2017.2677464.

Wang, B., Zheng, X., Qu, B. and Lu, X. (2020). Retrieval topic recurrent memory network for remote sensing image captioning, *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* **13**: 256–270. doi: 10.1109/JSTARS.2019.2959208.

Wu, S., Zhang, X., Wang, X., Li, C. and Jiao, L. (2020). Scene attention mechanism for remote sensing image caption generation, *2020 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–7. doi: 10.1109/IJCNN48605.2020.9207381.

Xu, N., Zhang, H., Liu, A., Nie, W., Su, Y., Nie, J. and Zhang, Y. (2020). Multi-level policy and reward-based deep reinforcement learning framework for image captioning, *IEEE Transactions on Multimedia* **22**(5): 1372–1383. doi: 10.1109/TMM.2019.2941820.

Yuan, Z., Li, X. and Wang, Q. (2020). Exploring multi-level attention and semantic relationship for remote sensing image captioning, *IEEE Access* **8**: 2608–2620. doi: 10.1109/ACCESS.2019.2962195.

Zhang, X., Wang, Q., Chen, S. and Li, X. (2019). Multi-scale cropping mechanism for remote sensing image captioning, *IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium*, pp. 10039–10042. doi: 10.1109/IGARSS.2019.8900503.

Zhang, X., Wang, X., Tang, X., Zhou, H. and Li, C. (2019). Description generation for remote sensing images using attribute attention mechanism, *Remote Sensing* **11**(6). doi: 10.3390/rs11060612.

Zhang, Z., Diao, W., Zhang, W., Yan, M., Gao, X. and Sun, X. (2019). Lam: Remote sensing image captioning with label-attention mechanism, *Remote Sensing* **11**(20). doi: 10.3390/rs11202349.

Zhou, D., Zhang, C., Li, Z. and Wang, Z. (2020). Multi-level visual fusion networks for image captioning, *2020 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8. doi: 10.1109/IJCNN48605.2020.9206932.