National College of Ireland

Bidirectional LSTM approach to image captioning with scene features

MSc Research Project
MSc. Data Analytics

Davis Munachimso Agughalam
Student ID: 19143354

School of Computing
National College of Ireland

Supervisors: Dr Paul Stynes
Dr Pramod Pathak

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Davis Munachimso Agughalam
Student ID:	19143354
Programme:	MSc. Data Analytics **Year:** 2019/2020
Module:	Research project
Supervisors:	Dr Paul Stynes and Dr Pramod Pathak
Submission Due Date:	17th August 2020
Project Title:	Bidirectional LSTM approach to image captioning with scene features
Word Count:	8147 **Page Count:** 19

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: Davis Agughalam

Date: 17th August 2020

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):	

Bidirectional LSTM approach to image captioning with scene features

Davis Agughalam
19143354

Abstract

Generating sentence descriptions for images is an area of research combining computer vision and natural language processing. More recently, it has been driven by encoder-decoder deep learning approaches where visual features are learned with a convolutional neural network (CNN) encoder are passed to a long short-term memory (LSTM) decoder for language generation. One major challenge in this approach is bridging the modality gap between the image and text data to enhance the semantic correctness of the generated sentences. While researchers have explored different features to achieve this, scene exploratory features have been largely underutilised and where utilised have been deployed with unidirectional LSTM decoders limited to retaining only past information thus producing poor results for long sequences. This research adopts a novel approach leveraging scene information deployed with a bidirectional LSTM decoder to achieve more semantically correct image descriptions. Pretrained CNNs Inceptionv3 and Places365 are employed for object and scene image feature extractions respectively before a bidirectional LSTM decoder is employed for language translation. This approach is validated by conducting experiments using the Flickr8k benchmark dataset and the results show improved performance compared to other encoder-decoder methods using just global image features thereby outlining the complementary advantages of scene information and bidirectional LSTMs to image captioning tasks.

1 Introduction

To generate captions for images, humans observe the details in the image scene to detect the objects in the scene and combine these objects to form semantically correct descriptions. For computers, this is challenging compared to related tasks including object detection and image classification where there has been huge success. This is because the generated sentences should contain words representing the underlying relationship between the objects in the image scene to be semantically correct to a reader thus cutting across both computer vision and natural language processing. The applications of such systems include screen readers and information retrieval for image search engines.

More recently, encoder-decoder deep learning techniques have dominated this domain of research phasing out others such as the template-based and retrieval-based approaches due to its flexibility, availability of required computing resources, superior performances and successes in other research domains and was pioneered by Vinyals et al. (2015). This approach has become state of the art in image captioning research entailing feature learning with a deep CNN suited for image processing as the encoder and a recurrent neural network

(RNN) best suited for machine translation tasks as the language translating decoder. Different variations to this approach have been employed by researchers aiming to make improvements such as the work done by Khamparia et al. (2019) where the long short-term memory (LSTM) variant of the RNN is used to overcome the exploding gradients problem faced by regular RNNs. Suggesting the need for extra features to augment the high level features extracted using CNNs, thus ably representing every aspect of the image especially in complex scenes, researchers have proposed different features such as semantic element embeddings, part of speech tags (POS) and object descriptors as was done by Zhang et al. (2019a). However, features directly identifying the image scene have not been employed exhaustively and have not been investigated at all with bidirectional LSTM decoders. Peng et al. (2019) investigated image scene features to improve generated sentence descriptions with unidirectional LSTMs retaining only past information for prediction sequences thus producing sub optimal results for longer sequences.

This research therefore extends this the work investigating image captioning using a novel framework adopting scene features and bidirectional LSTMs to improve the semantic richness of the generated captions. This thus raises the research question, to what extent can bidirectional LSTMs and scene exploratory features improve the semantic richness of automatically generated image descriptions? Experiments are conducted using the fairly limited Flickr8k benchmark dataset for image captioning research due to limited computing resources and results evaluated using the Bilingual evaluation understudy (BLEU) and recall-oriented understudy for gisting evaluation (ROUGE) metrics. The objectives developed to address the research question are further outlined as follows;

- Investigate state of the art research for image captioning and adopt best practices.
- Design a framework encompassing each section of the research implementation and proposed novelty.
- Implement the proposed CNN-BiLSTM approach while conducting experiments in stages for effective comparative analysis.
- Evaluate the performance of the approach using standard evaluation metrics for image captioning research such as BLEU and ROUGE scores.

Considering the research objectives, the major contribution of this research lies in the use of a combination of scene information and bidirectional LSTMs to generate semantically correct image captions in a novel framework targeted at improving state of the art approaches.

The rest of this research paper is presented as follows. Section 2 discusses related literature while section 3 describes the methodology in detail. Section 4 outlines the implementation of the approach, section 5 presents and discusses research results while section 6 presents a conclusion and discussion on the key findings.

2 Related Work

Image captioning as an area of research has garnered interest from various researchers over the years all aiming to improve existing approaches and generate semantically correct captions. Initially, retrieval and template-based approaches were mostly employed for image captioning tasks, but the advent of deep neural networks saw a shift from these previous approaches to deep neural work backed encoder-decoder approaches (Gupta and Jalal, 2019).

These approaches are reviewed in this section to identify best practices to guide the major decisions in this research and elaborate on the potential gaps in current state of the art.

2.1 Previous Image Captioning Approaches

Template and retrieval-based approaches were formerly mostly adopted for image captioning tasks. The template-based approaches typically involved object detection and language translation as independent subtasks (Liu et al., 2019). First, the objects in the image are detected using an object detection algorithm and these objects are expressed as sentences using different language models noting the attributes between the objects and the environment. A set number of predefined slots in a template are filled with obtained object relations and attributes to generate the final sentence description of images. This approach however, was rigid and generated sentence descriptions were not diverse (Liu et al., 2019). In retrieval-based methods, sentence descriptions for similar images in the training data are retrieved and transformed to represent the query image. This method required very large data to cover every possible query image thus is not scalable as retrieved captions are limited to only training images. These challenges in the use of template and retrieval methods fuelled the adoption of deep neural networks for end to end learning as proposed by Vinyals et al. (2015), for image captioning tasks as they are flexible and scalable.

2.2 Deep Neural Networks for Image Captioning

The encoder-decoder deep neural network framework when used for image captioning, have been shown to produce more flexible, natural and semantically richer sentence descriptions when compared to other approaches. Different enhancements such as employing semantic embeddings to enhance the semantic richness of the automatically generated captions and adopting the spatial or semantic attention mechanisms to bridge the gap between image and textual data modalities have been explored and are reviewed in detail herein.

Khamparia et al., (2019), Makav and Kilic (2019), Vinyals et al., (2015), and Kiros et al., (2014), employed a CNN-LSTM hybrid architecture for image caption generation. As is customary for encoder-decoder approaches, the images were processed for features using a pretrained CNN and these features are passed into the LSTM decoder for sentence generation. The LSTM unit ensured better performance than a regular RNN due to its ability to recall previous input data while obtaining new input and therefore being a better option for sequence to sequence and vector to sequence translations such as this. Their approach produced better results than template-based, and retrieval approaches further outlining the superiority of the encoder-decoder approach. However successful, only global image features were explored for the research as information from the text corpus was largely ignored.

2.2.1 Image captioning with Semantic Embeddings

Using extracted semantic information from the text corpus as part of the input features have also been explored as a means to improve the resulting sentence description of images. Zhang et al. (2019a) took this approach arguing that only CNN extracted visual features lack the ability to present scene complexity in the generated text translations. They employed semantic element information extracted from the text corpus such as part of speech tags (POS) as a means to bridge the gap in modality between the image and textual data and

consequently improve the generation of complex image descriptions. He et al. (2019) conducted similar research using POS tags as part of the input features in the LSTM. Considering other semantic features other than POS tags, Guo et al. (2019) used a structural set of attribute, object, scene and activity. Utilizing this set as part of the inputs into the language generation LSTM resulted in more semantically descriptive sentences. Similar research adopting semantic embeddings obtained through topic modelling with latent Dirichlet allocation (LDA) was conducted by Dash et al. (2019). Using LDA, they extracted topics for each of the captions in the text corpus and used these topics as inputs alongside the image features to guide the LSTM during sentence generation.

Other semantic information such as text information present in the image like signposts, billboards and labels can also be employed to improve the generated captions as most times the text in the image is directly present in the caption. Gupta and Jalal (2019) employed this approach extracting the textual cues using text saliency and spatial attention. Embeddings for the identified textual cues were learned and concatenated with the global image features and used as input for sentence generation in the LSTM. This approach produced detailed image captions when the textual cues spotted were relevant to the sentence description. However, the generated captions were incoherent when the textual cues were not related to the caption. The results of the research conducted by these researchers illustrated the advantages of leveraging information extracted from the text data to improve the generated captions as their approaches produced more semantically correct captions due to the use of information from text and image data modes.

Further leveraging embedded information to enhance the semantic content of automatic image descriptions, Peng et al. (2019) extracted scene information from the images and the text corpus. Latent Dirichlet Allocation (LDA) is used to extract scene information from the text corpus and two pretrained CNNs ResNet50 and Places365 CNN are used to extract the global and image scene features respectively. Instead of simple concatenation of features, a multilayer perceptron was trained to combine the image and text scene features to produce a scene vector. The global image features and the scene vectors were then fused into a double LSTM for subsequent sentence generation. Their results highlighted the positive impacts of scene features for image captioning tasks as there was huge improvement in the semantic richness of the captions. Their work was however limited by feature redundancy during word predictions as the entire image information is presented to the LSTM at each word interval. The work was also limited by the use of unidirectional LSTMs as they only retain past input information during model training while bidirectional LSTMs take cognizance of past and future inputs across the entire text sequence thereby providing better performance for longer text sequences.

2.2.2 Image captioning with Bidirectional LSTMs

To take advantage of both past and future information in the sequence during word translation and improve the results of the image captioning model, Wang et al. (2016) explored deep bidirectional LSTM architectures. They employed a CNN to extract the image features which were then passed into the deep Bi-LSTM architecture for word translation

while exploring various data augmentation techniques to counter overfitting in such deep networks. From their results, it is seen that the Bi-LSTM architecture improved generation of longer word sequences as its ability to generate consecutive words better was expressed in higher BLEU scores. Similar research conducted by Xiao et al. (2019a) using a bidirectional LSTM architecture in a dense semantic embedding network (DSEN) for image captioning illustrated the advantages of bidirectional LSTMs for longer sequences. Different architectures were tested for the DSEN using ResNet and VGG16 pretrained models and results also showed the advantages of Bi-LSTM network as their approach gave good scores for longer sequences.

2.2.3 Attention Based Image captioning

Using the entire global image features from the encoder in the decoder during model training is a recurring theme amongst the works reviewed above. This causes a waste of resources as most of the image features are not relevant during learning of particular words in the corpus during training and this negatively affects the results (Cao et al., 2019). Alternative approaches using the attention mechanism as part of the encoder-decoder framework have been adopted to solve this problem. Attention enhances the performance of deep neural networks for particular tasks as it amplifies relevant information in the data required for completing said tasks during model training. It mimics the human visual attention system as the brain selectively filters and amplifies information relevant to completing certain tasks for humans (Ding et al., 2019). For generating sentence descriptions for images, it is adapted to identify only image features relevant to generating words at each time step of the LSTM word generation sequence.

Dang et al. (2019) explored the importance of the attention mechanism in their work using two pretrained CNNs for multi-feature leaning. Comparing different architectures to test the effect of the attention mechanism, the results indicated that the attention mechanism improved performance significantly as the architecture with the attention layer performed better than the one without in terms of evaluation metrics. Similar research done by Zhang et al., (2019b) using the attention mechanism in an encoder-decoder architecture further buttressed its advantages. Xiao et al. (2019b) extended the attention mechanism using a double LSTM backed adaptive attention mechanism. This extension improved the visual attention mechanism as it used a visual sentinel to choose the time step and location to pay attention during word generation. Nezami et al. (2019) also adopted the attention mechanism aiming to enhance the emotional content of automatically generated descriptions by considering the facial expressions in the images. A facial expression recognition model (FER) was employed to detect emotions. To do this for more than one individual for images with several people, the attention mechanism was used to identify face representative features. These features used as input to the sentence generation LSTM alongside the overall image features to generate captions with meaningful emotional content as reflected in the BLEU scores.

Semantic attention in addition to visual attention has also been explored by researchers in an attempt to improve encoder-decoder architectures for image captioning. Liu et al. (2020) adopted dual attention using an InceptionV4 CNN for extracting the global image features and an LSTM network for generating the sentences. From their results, the use of both visual

and textual attention improved performance. Similar research conducted by Zhang et al. (2019c) for generating sentence descriptions in Chinese language also used a dual attention mechanism. The outputs of the individual attention modules were joined in a multimodal space and was adopted as input to an LSTM for sentence generation. The results also showed the advantages of employing the attention mechanism for image captioning. The research conducted by the above authors outline the importance of either visual or textual attention for image captioning as it also helps minimize the modality gap between the text and image features and selects relevant image regions during word prediction.

In conclusion, from the research reviewed herein, it is seen that semantic embeddings such as POS tags, text cues, word tetrads, and topics, bidirectional LSTM architectures and the attention mechanism have been employed to enhance the semantic richness of automatically generated image captions in encoder-decoder approaches by various researchers. However, isual embeddings have largely been underutilized. Peng et al. (2019) investigated the potential impact of image scene features as visual embeddings to improve semantic content of the generated captions but their work was limited as they used unidirectional LSTMs which are incapable of taking future information from the text sequences into consideration during the word translation sequence thus do not perform well for longer text sequences. Aiming to address this gap, this work investigates a novel image captioning approach leveraging deep image scene information as visual embeddings to enhance the semantic quality of the generated sentence descriptions while adopting bidirectional LSTMs to harness both past and future information during word predictions and consequently improve the overall performance and there does not appear to be any work that has adopted this combined approach for image captioning.

3 Research Methodology

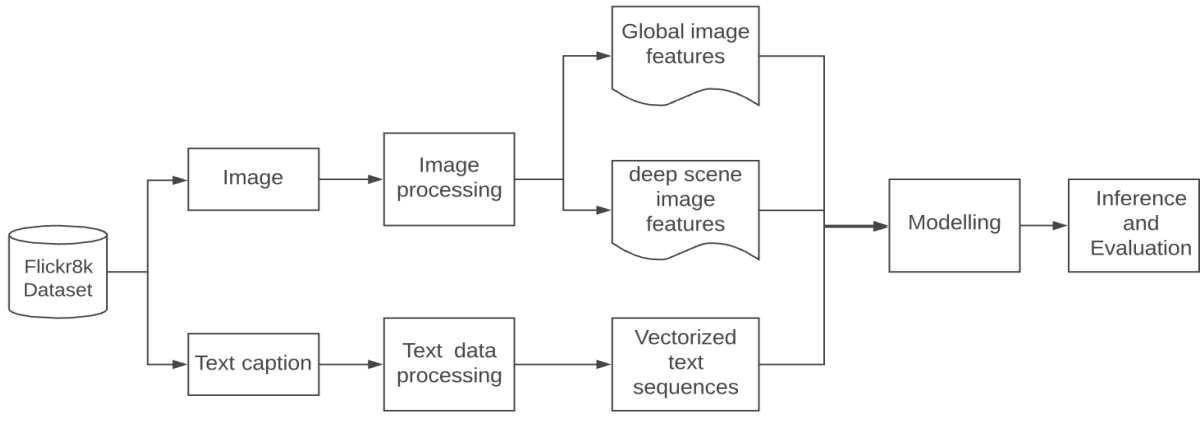

Figure 1: Approach methodology

The framework for the proposed approach is outlined in Fig. 1. Adopting the knowledge discovery in databases (KDD) system as is common in reviewed state-of-the-art research, the data is selected, processed, transformed and then used for modelling before final evaluation of results.

This section details the approach followed in carrying out this research from data acquisition and processing to the model design, specifications and configurations.

3.1 Dataset

The Flickr8k dataset is a benchmark dataset for image captioning research as established from reviewed literature (Peng et al., 2019, Vinyals et al., 2014, Kiros et al., 2014). Though there are other available datasets such as MSCOCO and Flickr30k, due to the limited computing resources, the Flickr8k dataset[1] is adopted for this research. The dataset contains 8000 images of human and animal activities with each having 5 ground truth sentence descriptions generated automatically using the Amazon mechanical Turk. The authors of the dataset already propose a standard split of 6000 for training, 1000 for development and 1000 for testing and the dataset is publicly available for research purposes.

3.2 Data Processing

To achieve satisfactory results, it is important to properly prepare and transform the data to the appropriate format. The processing steps for the sentence descriptions and images are outlined in the following sections.

3.2.1 Text Processing

- **Lower case conversion:** The dataset texts contain words with varying letter cases posing a problem to the model as same words with varying capitalisations would be regarded as different thereby increasing problem vocabulary and subsequently complexity. Hence, there is need to convert the entire text to lower case to avoid this.
- **Punctuation removal:** This research aims to produce descriptive sentences without punctuation for images hence the presence of punctuations add complexity to the problem which is beyond the scope of this work.
- **Number removal:** Numerical data present in the texts poses a challenge to the model as it increases vocabulary thus should be removed.
- **Indicate start and end sequence:** Word tokens ‘<start>’ and ‘<end>’ are added at beginning and end of each sentence to indicate the beginning and end of the prediction sequence to the model.
- **Tokenization:** The clean text is broken down into constituent words and a dictionary containing the entire vocabulary for word to index and index to word matchings is created.
- **Vectorisation:** The words in the text data are encoded using unique numerical representations from the word to index dictionary thus converting the cleaned sentence descriptions to numerical sequences before word representative vectors were learned from these sequences. To overcome varying sentence lengths, the shorter sentences are padded to the length of the longest sentence sequence.

[1] https://www.kaggle.com/shadabhussain/flickr8k

3.2.2 Image Processing and Feature extraction

The images are first loaded to an array before compression and normalisation to fit the input shape of the pretrained CNNs for feature extraction. CNNs are suitable models for extracting image features owing to their convolution and pooling processes and deeper CNN architectures are better at extracting image features (Chollet, 2017). In state-of-the-art research for image captioning such as the work done by Peng et al., (2019), due to the limited dataset, time required to train a model from scratch and complexity involved, a pretrained CNN with a deep architecture is used to extract the image features from the images rather than training from scratch. This research follows a similar approach using InceptionV3 and Places365 VGG16 pretrained models for global object and deep scene image features respectively.

The global object image features are focused on detecting the objects in the images and are extracted with the Inceptionv3 model. The model is trained on the ImageNet dataset with over 13 million images to recognize 1000 object classes hence can extract distinctive image features to detect these objects. This ensures better performance than a model trained from scratch on the limited research dataset and also saves training time.

The scene image features carry background knowledge indicating the context of the objects in the image in relation to each other and are complementary to the object features to generate more semantically correct captions. The VGG16 Places365 model for image scene recognition was also trained on a large dataset to identify scenery such as parks, restaurants and gyms in the images and was used to extract the image scene features in this research.

The features from the last fully connected layer of both networks are extracted as the image features.

3.3 Modelling

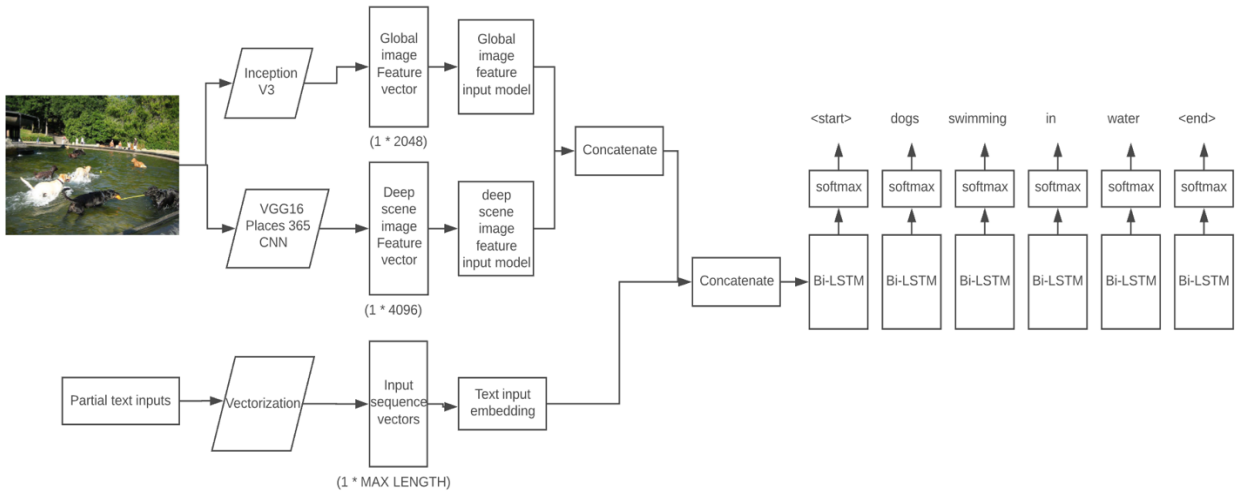

Figure 2: Encoder-decoder modelling

The encoder-decoder model approach is adopted in this research. The first part of the model encodes the image inputs while the second part of the model decodes it to text sequences. The model has 3 separate inputs namely, the global image features, the image scene features and the reference text. Two separate models consisting of input, dense and repeat vector layers to

repeat the output vectors across the entire length of the text sequence are built for each of the image inputs and their outputs concatenated. A text model consisting of an input and embedding layer learning the vector representations of the words in the text sequence is also built and the output combined with the concatenated image outputs in a separate model. The output from this concatenation is then passed to a fully connected Bi-LSTM layer before final connection to a dense layer with softmax activation predicting word probabilities.
To perform a valid comparative analysis, the experiments are conducted in stages. The first experiment is a model replicating state of the art encoder-decoder approaches using just the global image feature inputs and the text input for the modelling with a unidirectional LSTM. The second experiment takes the global image features, the deep scene image features and the text inputs with a unidirectional LSTM. To fully answer the research question, the last experiment is conducted using all image and text inputs modelled with a bidirectional LSTM.

3.4 Inference and Evaluation

Two approaches mostly applied for inference in neural machine language tasks are the greedy and beam search approach (Vinyals et al., 2015). For the greedy search approach, the word with the highest softmax probability is chosen at each word prediction step while the beam search method considers more probabilities using a pre-set beam length. A cumulative probability for all the likely words is obtained and the words with the highest probabilities are selected as inference results. As was seen in Vinyals et al. (2015), the greedy approach performs better than the beam search for inference purposes hence is the adopted approach in this research.
As image captioning varies from other traditional machine learning problems, common metrics such as accuracy and precision are not sufficient to evaluate generated sentences. Specialised evaluation metrics such as the Bilingual evaluation understudy (BLEU) and recall-oriented understudy for gisting evaluation (ROUGE) have been developed and have served as benchmarks for evaluating image captioning results (Khamparia et al., 2019, Hossain et al., 2018). The BLEU is a precision backed metric that compares individual words of the generated sentences with words in the reference text and provides scores depending on the average number of matches. The BLEU metric is limited as it does not evaluate long sentences efficiently (Hossain et al., 2018) and offers no sense of semantic understanding thus other metrics such as the ROUGE were developed. The ROUGE system matches word pairs, n-grams and sequences from the generated text to matches in the reference corpus and returns values based on the number of co-occurrences. The ROUGE-L variant of the ROUGE metrics have been mostly used to evaluate image captioning and returns scores based on the longest common sequence of words between the generated and reference text corpus (Liu et al. 2019). Both metrics are used to evaluate the approaches in this research.

4 Design Specification

The underlying algorithms powering this approach are discussed in detail in this section. Convolutional neural networks and an LSTM are the core networks in this research for image feature encodings and language translation respectively. These models are discussed in detail as follows.

4.1 Convolutional neural network (CNN)

A convolutional neural network (CNN) is a popular type of neural network suitable for processing and understanding image features through a series of convolution and pooling layers (Chollet, 2017). The network performs a convolution operation in the convolution layer extracting features from the input image.

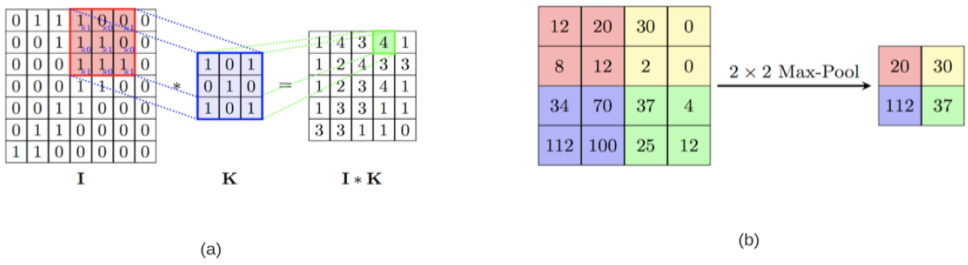

Figure 3: Convolution (a) and Max-pooling (b) processes[2]

The convolution process involves sliding a filter of pre-set size across the image input matrix at a pre-set slide to produce feature maps and to overcome nonlinearity in this process, a rectified linear unit activation function is adopted (Chollet, 2017). The overall purpose of the convolution operation is to extract the high-level features from the axes and channels of the image. After convolution, a pooling operation is done to reduce the spatial size of the high-level features obtained. The most distinctive features wholly representing the entire image are identified after pooling. After a series of convolution and pooling processes, the resulting feature vectors are flattened before being put through a fully connected network (FCN) to obtain probabilities for the multiple prediction classes. The layers in these steps are trained end to end using backpropagation. This combination of processes account for the network's ability to carry out complex image processing tasks. The InceptionV3 and VGG16 networks follow this architecture thus can extract highly distinctive image features as they are trained on very large datasets.

4.2 Long Short-Term Memory Network (LSTM)

The long short-term memory (LSTM) extension of the recurrent neural network (RNN) has been most adopted for image captioning tasks as it overcomes the vanishing gradient and short-term memory issues faced by an ordinary RNN during backpropagation especially for tasks involving long sequences (Cao et al., 2019). The LSTM contains a cell unit which functions as a memory block that can enhance the network's memory capacity and control flow of information with different gate categories. The unit consists of the input gate, the forget gate and the output gate. The input gate determines which information from the previous steps is relevant to the current step, the forget gate determines what part of the input information should be discarded and the output gate determines the next hidden state

[2] http://people.mines-paristech.fr/fabien.moutarde/ES_MachineLearning/Practical_deepLearning-convNets/convnet-notebook.html

(Chollet, 2017). This controlled flow of information and ability to recall past inputs effectively makes the LSTM network an optimal choice for image captioning tasks. While unidirectional LSTMs incorporate past input information in the current step, bidirectional LSTMs better control information flow by being exposed to both past and future sequences during learning hence is a better choice for more accurate predictions though more computationally expensive.

5 Implementation

This section presents detailed information on the implementation of the proposed approach in this research.

5.1 Setup

The deep neural networks in this research are implemented using the Google Collaboratory cloud environment. This is due to the availability of a graphics processing unit (GPU) on the platform. GPUs are able to train neural networks faster than CPUs due to its ability to perform parallel mathematical computations on multiple batches of data at the same time while CPUs are limited in task concurrency.

The host machine used to implement this research is a MacBook Pro with a core i7 processor, 16GB RAM and an AMD 5600 GPU. The Jupyter notebook which is part of the anaconda distribution package is used for the implementation of this work both on the local machine and the cloud environment.

5.2 Data Handling

Due to the size of the dataset, data processing was carried out on the host machine and the output was stored as pickle files before transferring onto the cloud environment for model training. The dataset was processed using custom functions written in python to carry out various pre-processing steps. Figure 2 below shows a sample image and its text descriptions after text pre-processing. Due to the varying domains from which the images in the dataset come from and to ensure better model generalisation, the vocabulary was not capped and a vocabulary size of 8862 was realized after processing.

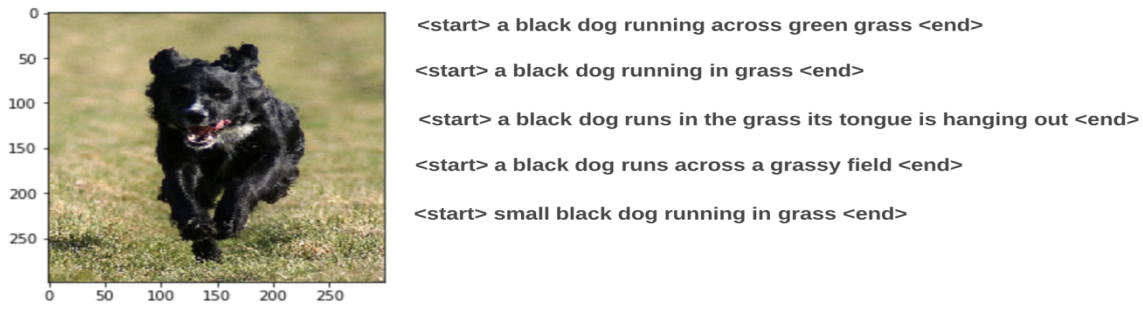

Figure 4: Sample normalized image and processed text descriptions

After text cleaning, the resulting texts are broken down into constituent words and word to index and index to word matching dictionaries are created using the indexes of the words in the python list. These matching dictionaries are used to convert the text sequences into vector

sequences of length 37 corresponding to the length of the longest text sequence. Shorter vector sequences are padded to fit this length.

For the images, the global image feature vectors are extracted from the last fully connected layer of the InceptionV3 pretrained CNN from Keras[3] applications. Before extraction, the images are first compressed to 299x299 arrays to fit the Inceptionv3 model input and the pixels normalised to a scale of 0 to 255. For the image scene features, a VGG16 Places365 pretrained model is used. The images are compressed to 224x224 to fit the input shape of the model and also normalised on a scale of 0 to 255. The output feature vectors from both networks are stored in pickle files for easy upload to the cloud platform for modelling.

5.3 Modelling

Three different models were built in total using this base architecture with varying number of inputs and layer types to conduct different experiments.

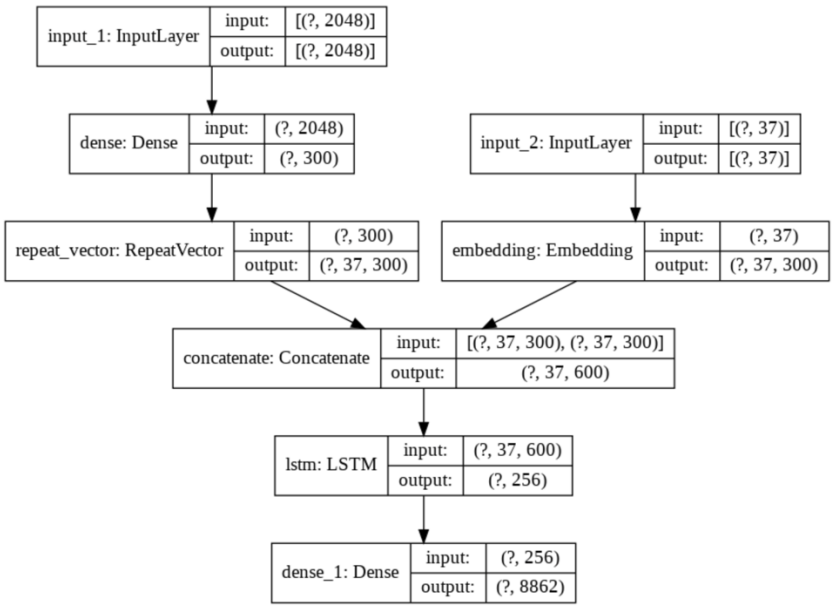

Figure 5: Baseline CNN-LSTM model architecture

The layers are described as follows;

- **Image model:** This is the first layer of the model that takes the encoded image feature vectors. The input shape is dependent on resulting feature set from the image encoding model. For the global image features extracted using the InceptionV3, the input shape was 2048, and the scene features extracted using the VGG16 Places365 CNN had 4096 as feature dimensions thus the input shape for its model corresponded to this. This is followed by a dense layer that reduces the input dimension to match with the dimension of the text embeddings set at 300. A repeat vector layer is used to repeat the outputs from the dense layer for the max length of the text descriptions.
- **Text model:** The text input model is responsible for taking the processed text sequences. Its input shape is the maximum length of the train sequences which is 37

[3] Keras.io

as all the sentence descriptions are padded to this length. After the input layer, an embedding layer to learn the vector representation of the input sequences follows. The embedding layer learns vector representations of each of the words in the text sequence to avoid matrix sparsity that can occur in other forms of word representations such as one hot encoding. An embedding dimension of 300 is used causing each word to be represented by a 300 dimension vector in the output.

- **Combined Model:** The outputs from the image model and the text model are concatenated and passed to another LSTM layer with 256 units. The variation for the different experiments are done in this layer as different experiments are conducted using unidirectional LSTM and bidirectional LSTM layers. The final layer of the model is a dense layer with the vocabulary size of 8862 as the number of units and softmax activation to predict the likelihood of each word to be the next in the sequence.

The layers are stacked using the keras functional API as described in the Fig. 5 for the first experiment with variations made for each subsequent experiment. The architecture of each of the models is described in detail in the configuration manual.

5.4 Training

The models are trained on 6000 images and validated and tested on 1000 images each. For optimum use of computing resources, the data is progressively loaded using a data generator to deliver data in batches of 128 to the model to avoid loading all images into RAM at the same time. The above described model architectures are compiled using 'categorical_crossentropy' as the loss and 'RMSprop' as the backpropagation optimizer. This is a popular choice in most of the observed literature due to its efficiency and faster convergence for neural language translation tasks (Vinyals et al., 2014, Wang et al., 2016). To avoid overfitting, early stopping was used. The model was also checkpointed, and weights saved as validation loss decreased. Each epoch took approximately 11 minutes to run on Google Colab. As the models had different capacities, the number of epochs required to reach full predicting potential varied so experiments were conducted testing the networks at 5, 10 and 15 epochs each. This also helped control overfitting.

6 Results and Evaluation

This section presents the findings of the three experiments conducted in this research. The first experiment is a replication state of the art, the second is an extension of the state of the art using scene features and the third experiment involves further extension of the state of the art using both scene features and a bidirectional LSTM. Each of these experiments were evaluated with BLEU and ROUGE metrics across the different number of training epochs due to their different model parameter capacities.

6.1 Experiment 1: CNN-LSTM model

To carry out effective comparative analysis and accurately measure the impact of the scene features and use of a bidirectional LSTM, an experiment is conducted without them using just

the global image features and a unidirectional LSTM. The model is evaluated at different number of training epochs to monitor overfitting and results are presented in Table 1.

Table 1: CNN-LSTM model BLEU and ROUGE Scores

Epochs	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-L
5	46.7	28.5	19.4	9	25.7
10	53.1	32.9	22.6	10.8	26.9
15	49.7	31.2	21.9	11.3	26.8

It is seen that the model starts overfitting after 10 epochs as the performance begins to decline. After 10 epochs, the model performance reduced from a BLEU-1 score of 53.1 to 49.7 at 15 epochs. The ROUGE scores also dropped from 26.9 at 10 epochs to 26.8 indicating deteriorating performance and overfitting.

Aiming to improve the performance of this model, scene features are introduced in the next experiment.

6.2 Experiment 2: CNN-LSTM model with scene factors

This experiment extends the previous CNN-LSTM model with an addition of scene features in a dual CNN feature approach. From the results shown in Table 2, a similar pattern to the CNN-LSTM experiment is noticed as the training performance also peaked at 10 epochs of training with a BLEU-1 score of 51.8 and ROUGE score of 26.67. The results also indicate that the addition of image scene features caused a reduction in the performance of the model as BLEU-1 scores dropped from 53.1 as shown in Table 1 to 51.8. This may be due to the need for increased model capacity to learn as new features were introduced to the model.

Table 2: CNN-LSTM with image scene factors BLEU and ROUGE Scores

Epochs	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-L
5	49.1	30.1	20.7	10.0	26.4
10	51.8	31.6	20.8	9.8	26.67
15	45.8	27.7	18.6	8.8	25.3

To curb this perceived limitation, a further experiment is conducted by increasing the capacity of the network using Bi-LSTMs.

6.3 Experiment 3: CNN-BiLSTM model with scene factors

In this experiment, the capacity of the model is increased as bidirectional LSTMs are used instead of unidirectional LSTMs.

Table 3: CNN-BiLSTM with image scene factors BLEU and ROUGE Scores

Epochs	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-L
5	55	34.9	24.8	13.1	27.9
10	53.4	33.7	24.0	12.4	27.4
15	46.8	29.5	20.9	10.8	26.0

From the results shown in Table 3, the performance of the model increased and peaked at 5 epochs of training. This model trained faster than the others but showed an increasing ability to overfit as performance started dropping after 5 epochs from a BLEU-1 score of 55 to 53.4 at 10 epochs. Even with declining performance due to overfitting, the results after 10 epochs of training still showed better performance than the CNN-LSTM model with scene factors deployed with unidirectional LSTMs in the previous experiment with a BLEU-1 score of 53.4 compared to 51.8. This further buttress the advantage of introducing a Bi-LSTM.
This CNN-BiLSTM with scene features experiment showed higher BLEU and ROUGE scores compared to the best reported values for the previous experiments thus having the overall best performance.

6.4 Discussion

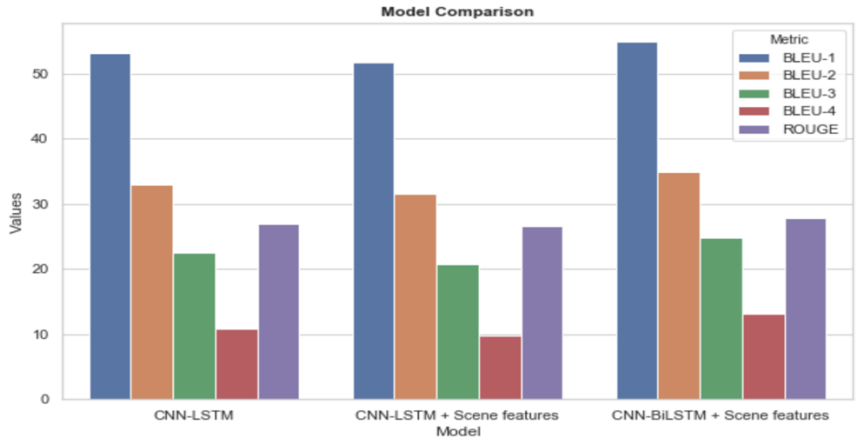

Figure 6: Experiment results comparisons

The results of the experiments conducted in this research show the ability of scene factors and bidirectional LSTMs to improve the performance of image captioning models.
The first experiment is conducted with just the CNN extracted global object image features and text vectors as inputs for effective comparative analysis with subsequent experiments. To monitor overfitting due to limited data, the models are evaluated at 5, 10 and 15 epochs of training. It is seen that the first model peaks at 10 epochs with 53.1 BLEU-1 score and performance goes downhill from there as indicated by BLEU-1 OF 49.7 at 10 epochs.
Using this as a base to accurately ascertain the effects of the scene features on the performance of the model, the next experiment introduces the image scene factors as part of the model training features. Under the same testing conditions, the model peaks at 10 epochs also but does not perform as well as the model without the scene features. There is an indication that the newly introduced features inhibited the model training process as shown in the maximum BLEU-1 score of 51.8 compared to the previous experiment BLEU-1 score of 53.1. This spurred the need for further analysis and experimentation.
Looking to resolve the limitation of the previous experiment, the third experiment introduces a Bi-LSTM in the model architecture to improve the model capacity and take advantage of both past and future information in the text sequences while training. Following this introduction, the performance peaks at 5 epochs showing that the model trains and converges faster while performing better as indicated in a BLEU-1 score of 55. It is also seen that the bidirectional LSTMs and scene factors also improved the semantic meaning of the sentences

for longer sequences as was shown in improved BLEU-2, BLEU-3 and BLEU-4 scores. The importance of the Bi-LSTM is further buttressed as even with declining performance at 10 epochs, it still outperformed the experiment with scene factors and unidirectional LSTMs. However, the model starts overfitting early as performance declines after 5 epochs. This early overfitting thus implies that while improving the model capacity is advantageous, care has to be taken to identify optimum hyperparameter settings to ensure maximum performance while being robust to overfitting.

Figure 7: Models and generated descriptions

Further qualitative analysis is conducted to ascertain the performance of the algorithms and qualitatively evaluate the results of some of the predictions are shown in Fig. 7. The model with scene information and a bidirectional LSTM is able to appropriately capture the background information in the image thus producing sentences with more semantic meaning. In the first image, the advantages of adopting scene features was expressed as the model was able to recognise a ‘crowd’ and introducing the Bi-LSTM network created a sentence with more complete semantic meaning identifying ‘a group of people’ and ‘a crowd’ as is the case in the image. For the second image, the model with scene features and a Bi-LSTM was once again able to correctly identify and apply relevant scenery as seen in the words ‘grassy field’ in the sentence description. Similar understanding and application of relevant scenery to generate a semantically rich caption was demonstrated by the model with scene features and Bi-LSTM again for the third image as compared to the first and second models as the words ‘rocky wall’ were introduced thus making the sentence more semantically correct at describing the activities going on in the image.

From the results discussed herein, it is seen that scene features are complementary to the global image features as it helps improve the generated captions and this is further the case when deployed with bidirectional LSTMs.
Compared to work done by Peng et al. (2019) using double unidirectional LSTMs, this approach saw an improvement in BLEU-3 and BLEU-4 scores for the adopted bidirectional LSTMs indicating superior performance for longer sequences. This further validates the hypothesis that bidirectional LSTMs can be leveraged to improve word translations for longer sequences.

7 Conclusion and Future Work

Aiming to produce semantically richer captions for images, this research adopts a novel approach investigating to what extent a combination of image scene features and bidirectional LSTMs in an encoder-decoder architecture can improve the semantic meaning of automatically generated image captions image. As part of the outlined objectives, an approach framework is designed, and experiments are conducted using the Flickr8k benchmark image captioning dataset.
The main finding of this research is the indication that the proposed approach using scene features and bidirectional LSTMs improved the generated captions as was evident from the results thereby answering the research question. This can be attributed to the ability of scene features to capture background context between the objects in the image. The use of bidirectional LSTMs also impacted the results positively as they are able to take into consideration past and future information during the word translation sequence. It is also found out that introducing scene features using unidirectional LSTMs inhibits the performance thus the need to improve model learning capacity by introducing the Bi-LSTM approach. This consequently improves performance and causes faster model convergence though it increases the likelihood of overfitting due to a small data size. Comparing the approach with previous approaches in image captioning, there is an increase in semantic correctness of the generated captions for longer sequences as indicated by the evaluation metrics.
Taking note of overfitting as a problem due to dataset size, the experiments were monitored by testing at different epochs. This is a limitation in this research as all possible number of epochs were not tested for and hyperparameter optimisation to determine the hyperparameters for optimum performance was not pursued due to limited computing resources.
For future work, this research can be extended using the attention mechanism to bridge the gap between the image and text modalities by mapping the texts to specific image regions during word translation to lead to further increase in semantic correctness of the generated captions. Evaluation metrics able to compare reference and generated captions based on semantic meaning and not just based on n-gram matches can also be developed and adopted for more complete evaluation.

Acknowledgement

I would like to use this opportunity to thank everyone who helped me achieve this final project. I am very thankful to my supervisors Dr Paul Stynes and Dr Pramod Pathak for their regular guidance and feedback throughout the course of this research project. I would also like to thank Dr Sachin Sharma for his guidance through the prerequisite course for this module and setting the stage for this project. Finally, I would like to specially thank my family, without whom none of these would be possible for their support throughout the entire period of my research.

References

Cao, D., Zhu, M. and Gao, L. (2019) 'An image caption method based on object detection', *Multimedia Tools and Applications*, 78(24), pp. 35329-35350. SpringerLink. doi: 10.1007/s11042-019-08116-9.

Chollet, F. (2017) *Deep learning with python.* 1st edn. Manning Publications Co.

Dang, T.X., Oh, A., Na, I.S. and Kim, S.H. (2019), 'The Role of Attention Mechanism and Multi-Feature in Image Captioning', in *Proceedings of the 3rd International Conference on Machine Learning and Soft Computing (ICMLSC).* Da Lat, Vietnam, January 2019, pp. 170-174, ACM Digital Library. doi: 10.1145/3310986.3311002.

Dash, S.K., Acharya, S., Pakray, P., Das, R. and Gelbukh, A. (2019) 'Topic-Based Image Caption Generation', *Arabian Journal for Science and Engineering*, 13(4), pp.1-10. SpringerLink. doi: 10.1007/s13369-019-04262-2.

Ding, S., Qu, S., Xi, Y., Sangaiah, A.K. and Wan, S. (2019) 'Image caption generation with high-level image features', *Pattern Recognition Letters*, 123(2), pp. 89-95. ScienceDirect. doi: 10.1016/j.patrec.2019.03.021.

Guo, R., Ma, S. and Han, Y. (2019) 'Image captioning: from structural tetrad to translated sentences', *Multimedia Tools and Applications*, 78(17), pp. 24321-24346. SpringerLink. doi: 10.1007/s11042-018-7118-7.

Gupta, N. and Jalal, A.S. (2019) 'Integration of textual cues for fine-grained image captioning using deep CNN and LSTM', *Neural Computing and Applications*, 48(3), pp. 1-10. SpringerLink. doi: 10.1007/s00521-019-04515-z.

He, X., Shi, B., Bai, X., Xia, G.S., Zhang, Z. and Dong, W. (2019) 'Image caption generation with part of speech guidance', *Pattern Recognition Letters*, 119(4), pp. 229-237. ScienceDirect. doi: 10.1016/j.patrec.2017.10.018.

Hossain, M.Z., Sohel, F., Shiratuddin, M.F. and Laga, H. (2019) 'A comprehensive survey of deep learning for image captioning', *ACM Computing Surveys (CSUR)*, 51(6), pp.1-36. ACM Digital Library. doi: doi.org/10.1145/3295748.

Khamparia, A., Pandey, B., Tiwari, S., Gupta, D., Khanna, A. and Rodrigues, J.J. (2019) 'An Integrated Hybrid CNN–RNN Model for Visual Description and Generation of Captions', *Circuits, Systems, and Signal Processing*, 39(2), pp.1-13. SpringerLink. doi: 10.1007/s00034-019-01306-8.

Kiros, R., Salakhutdinov, R. and Zemel, R.S. (2014) 'Unifying visual-semantic embeddings with multimodal neural language models', *arXiv preprint arXiv:1411.2539*, pp. 1-13. Available at: https://arxiv.org/pdf/1411.2539.pdf [Accessed: 3 Mar 2020]

Liu, X., Xu, Q. and Wang, N. (2019) 'A survey on deep neural network-based image captioning', *The Visual Computer*, 35(3), pp. 445-470. SpringerLink. doi: 10.1007/s00371-018-1566-y.

Liu, M., Li, L., Hu, H., Guan, W. and Tian, J. (2020) 'Image caption generation with dual attention mechanism', *Information Processing & Management*, 57(2), p. 102178. ScienceDirect. doi: 10.1016/j.ipm.2019.102178.

Makav, B. and Kılıç, V. (2019), 'Smartphone-based Image Captioning for Visually and Hearing Impaired', in *2019 11th International Conference on Electrical and Electronics Engineering (ELECO).* Bursa, Turkey, Turkey, 28-30 November 2019, pp. 950-953, IEEE Xplore. doi:10.23919/ELECO47770.2019.8990395.

Nezami, O.M., Dras, M., Wan, S. and Paris, C. (2020) 'Image captioning using facial expression and attention', *arXiv preprint,* arXiv:1908.02923, pp. 1-28. Available at: https://arxiv.org/pdf/1908.02923.pdf [Accessed: 3 Mar 2020].

Peng, Y., Liu, X., Wang, W., Zhao, X. and Wei, M. (2019) 'Image caption model of double LSTM with scene factors', *Image and Vision Computing*, 86(3), pp. 38-44. ScienceDirect. doi: 10.1016/j.imavis.2019.03.003.

Vinyals, O., Toshev, A., Bengio, S. and Erhan, D. (2015) 'Show and tell: A neural image caption generator', in 2015 *Proceedings of the IEEE conference on computer vision and pattern recognition.* Boston, MA, USA, 7-12 June 2015, pp. 3156-3164, IEEE Xplore. doi: 10.1109/CVPR.2015.7298935.

Wang, C., Yang, H., Bartz, C. and Meinel, C. (2016) 'Image captioning with deep bidirectional LSTMs', in *Proceedings of the 24th ACM international conference on Multimedia*, New York, NY, USA, October 2016, pp. 988-997, ACM Digital Library. doi: 10.1145/2964284.2964299.

Xiao, F., Gong, X., Zhang, Y., Shen, Y., Li, J. and Gao, X. (2019) 'DAA: Dual LSTMs with adaptive attention for image captioning', *Neurocomputing*, 364(5), pp.322-329. ScienceDirect. doi: 10.1016/j.neucom.2019.06.085.

Xiao, X., Wang, L., Ding, K., Xiang, S. and Pan, C. (2019) 'Dense semantic embedding network for image captioning', *Pattern Recognition*, *90*, pp.285-296. ScienceDirect. doi: 10.1016/j.patcog.2019.01.028

Zhang, X., He, S., Song, X., Lau, R.W., Jiao, J. and Ye, Q. (2019) 'Image captioning via semantic element embedding', *Neurocomputing*. [In press]. ScienceDirect. doi: 10.1016/j.neucom.2018.02.112.

Zhang, X., Tang, X., Wang, X., Zhou, H. and Li, C. (2019) 'Description generation for remote sensing images using attribute attention mechanism', *Remote Sensing*, 11(6), p. 612-630. Available at: https://www.mdpi.com/2072-4292/11/6/612 [Accessed: 18 Mar. 2020].

Zhang, Y. and Zhang, J. (2019) 'Application of Dual Attention Mechanism in Chinese Image Captioning', *Journal of Intelligent Learning Systems and Applications*, 12(1), pp.14-29. Scientific Research Publishing. doi: 10.4236/jilsa.2020.121002.