Major Project Final Report

On

Image Captioning with Region-Based Attention and Scene-Specific Contexts

Submitted by

15IT111 Viranchi Badheka
15IT132 Nihar Raichada
15IT231 Ria Kulshrestha
15IT243 Shivani Shrivastava

Under the Guidance of

Mr. Dinesh Naik
Assistant Professor

in partial fulfillment for the award of the degree of

Bachelor of Technology
in
Information Technology

Date of Submission: April, 2019

Department of Information Technology
National Institute of Technology Karnataka, Surathkal
2018-2019

Department of Information Technology
National Institute of Technology Karnataka, Surathkal

Certificate

This is to certify that the project entitled, **Image Captioning with Region-Based Attention and Scene-Specific Contexts** submitted by Viranchi Badheka, Nihar Raichada, Ria Kulshrestha and Shivani Shrivastava, in partial fulfillment of the requirements for the award of B.Tech in **Information Technology** at the **National Institute of Technology Karnataka, Surathkal** is an authentic work carried out by them under my supervision and guidance.

1. **Viranchi Badheka** **15IT111**
2. **Nihar Raichada** **15IT132**
3. **Ria Kulshrestha** **15IT231**
4. **Shivani Shrivastava** **15IT243**

Mr. Dinesh Naik
Project Guide, Dept of IT
NITK Surathkal, Mangalore

Prof. G. Ram Mohana Reddy
Chairman-DUGC, Dept of IT
NITK Surathkal, Mangalore

Department of Information Technology
National Institute of Technology Karnataka, Surathkal

Declaration

We, hereby declare that the project entitled, **Image Captioning with Region-Based Attention and Scene-Specific Contexts** is being submitted to National Institute of Technology Karnataka during the academic year 2018 2019 for the award of the degree of Bachelor of Technology in **Information Technology**, is a bonafide report of the work carried out by us. The material content in this project work report has not been submitted in any other university or institution for the award of any degree.

Name of Student	Register No.	Signature with Date
1. **Viranchi Badheka**	**15IT111**	
2. **Nihar Raichada**	**15IT132**	
3. **Ria Kulshrestha**	**15IT231**	
4. **Shivani Shrivastava**	**15IT243**	

Place:

Date:

Abstract

A region based attention model that automatically learns to describe the content of images using Recurrent Neural Networks and Scene-Specific contexts. The image captioning system will exploit the parallel transitioning dynamics between the visual focus and the sentences. Scene-specific contexts will capture higher-level semantic information encoded in an image. The contexts will adapt to language models for word generation to specific scene types The performance of the model is bench marked on the MS COCO dataset. Deep network attention can be viewed as a form of alignment from language space to image space. These attention maps carry important information in understanding deep networks.

Keywords: Image captioning, visual attention, scene-specific context, LSTM.

Contents

List of Figures

List of Abbreviations

LDA	Latent Dirichlet Allocation
RNN	Recurrent Neural Network
RCNN	Region Convolutional Neural Network
SVM	Support Vector Machine
ROI	Region of Interest
LSTM	Long Short Term Memory
FRCNN	Fast Region Convolutional Neural Network

1 Introduction

Despite the recent progress in automatic generation of image captions, it remains a challenging task. The existing systems have attained very promising results by leveraging several crucial advances in computer vision and machine learning: optimizing on datasets having large number of images and their corresponding human-annotated captions and learning complex models that are capable of generating human-readable sentences.

1.1 Motivation

Automatically generating captions of an image is a task very close to the heart of scene understanding and is one of the primary goals of computer vision. Not only must caption generation models be powerful enough to solve the computer vision challenges of determining which objects are in an image, but they must also be capable of capturing and expressing their relationships in a natural language. For this reason, caption generation has long been viewed as a difficult problem. It is a very important challenge for machine learning algorithms, as it amounts to mimicking the remarkable human ability to compress huge amounts of salient visual information into descriptive language.

One of the most curious facets of the human visual system is the presence of attention. Rather than compress an entire image into a static representation, attention allows for salient features to dynamically come to the forefront as needed. This is especially important when there is a lot of clutter in an image. One advantage of including attention is the ability to visualize what the model sees. We implemented a model that can attend to salient part of an image while generating its caption.

2 Literature Review

2.1 Literature Survey

Image caption generation has long been a challenging problem in computer vision. Recent work aim to automatically generate words with language models learn from data.

Automatically describing the content of an image is a fundamental problem that connects computer vision and natural language processing. The state of the art architectures used in both are combined in [1] which proposed the idea of a neural and probabilistic framework to generate descriptions from images. They made use of a recurrent neural network(RNN) which encodes the variable length input into a fixed dimensional vector, and uses this representation to decode it to the desired output sentence.
But they replaced the encoder RNN by a deep convolution neural network (CNN). As CNNs can produce a rich representation of the input image by embedding it to a fixed-length vector.

There is an inherent limitation in the above approach because of the fixed length vector implying a static nature in cation lengths. Another more implementation based drawback is pointed out by [2] which reveals a significant limitation of simple RNN models which strictly integrate state information over time known as the vanishing gradient or exploding gradient problem which makes it really hard to learn and tune the parameters of the earlier layers in the network and worsens as the number of layers in the architecture increases.
LSTMs solve this by incorporating memory units that explicitly allow the network to learn when to forget previous hidden states and when to update hidden states given new information.
They proposed adding additional depth to LSTMs by stacking them on top of each

other and using the hidden state of the LSTM in current layer as the input to the LSTM in next layer.
They show stacked LSTM models provide improved recognition on conventional video activity challenges and enable a novel end-to-end optimizable mapping from image pixels to sentences.
They use an multimodal encoder-decoder model based on LSTM networks. They used a visual conv-net to encode a deep state vector, and an LSTM to decode the vector into a natural language string.

In order to fuse text information and visual features extracted on the whole image [3] proposed a multi-modal Recurrent Neural Networks (m-RNN) model to address both the task of generating novel sentences descriptions for images, and the task of image and sentence retrieval. The whole m-RNN architecture contains a language model part which learns the dense feature embedding for each word in the dictionary and stores the semantic temporal context in recurrent layers, an image part which contains a deep Convulutional Neural Network (CNN) that extracts image features and a multimodal part which connects the language model and the deep CNN together by a one-layer representation.

One of the most curious facets of the human visual system is the presence of attention (Rensink, 2000; Corbetta Shulman, 2002). Rather than compressing an entire image into a static representation, attention allows for salient features to dynamically come to the forefront as needed, [4] approaches to caption generation by incorporating the attention mechanism. Rather than compressing an entire image into a static representation, attention mechanism allows for salient features to dynamically come to the forefront as needed. This is useful when there is a lot of clutter in an image.
The paper proposes the Encoder-Decoder framework.Encoder is responsible for extraction of visual features from corresponding 2D portions of the images.The

extraction is done from lower convolutional layers that allows selective focus on certain parts of an image by selecting a subset of all the feature vectors.
Decoder uses LSTM that produces a caption by generating one word at every time step conditioned on a context vector, the previous hidden state and the previously generated words. For each location 'i' extracted by encoder, the mechanism generates a positive weight that represents the probability that location 'i' is the right place to focus for producing the next word, hence finding the region of interests.

The visual attention mechanism also plays an important role in natural language descriptions of images biased towards semantics. In particular, people do not describe everything in an image. Instead, they tend to talk more about semantically more important regions and objects in an image. [5] proposes the algorithm that combines the bottom-up and top-down approaches of image captioning through a semantic attention model.
Their definition for semantic attention in image captioning is the ability to provide a detailed, coherent description of semantically important objects that are needed exactly when they are needed.The semantic attention model attends a semantically Region of Interest(ROI) in an image,weighs the relative strength of attention paid on multiple ROIs and is able to switch attention among ROIs dynamically.
The top-down paradigm starts from a visual feature representation of an image and converts it into words, The bottom-up paradigm shows the changes of the attention weights for several candidate ROIs with respect to the recurrent neural network iterations.
The model is built on top of a Recurrent Neural Network (RNN), whose initial state captures global information from the top-down feature. As the RNN state transits, it gets feedback and interaction from the bottom-up attributes via attention mechanism.
The feedback allows more accurate and robust inferences of semantic gap between generated caption and actual image content.

The total number of images and windows to evaluate in an exhaustive search is huge and growing, it is necessary to constrain the computation per location and the number of locations considered.[6] proposes "Selective Search" by using segmentation to generate a limited set of locations(Regions of Interest), i.e. best locations for object recognition. they put most emphasis on recall and good object approximations over exact object boundaries. They used a full segmentation hierarchy and accounted for as many different scene conditions as possible, such as shadows, shading, and highlights, by using a variety of invariant colour spaces. They started with oversegmentation the image, i.e. a set of small regions which do not spread over multiple objects. Starting from the initial regions, they used a greedy algorithm which iteratively groups the two most similar regions together and calculates the similarities between this new region and its neighbours. This process is continued until the entire image becomes a single region.
Either all segments throughout the hierarchy (including initial segments) or the tight bounding boxes around these segments are considered as potential object locations.
The similarity between any two regions is defined as a sum of the fraction of the image that the segment a and b jointly occupy and defined as the histogram intersection between SIFT-like texture measurements. Both the factors are equally weighted.

It is of significance to improve the quality of candidate bounding boxes and to take a deep architecture to extract high-level features. To solve these problems, [7] proposed a model called "RCNN - Region with CNN" which adopts selective search [6] to generate about 2k region proposals for each image.
Their system consists of three modules. The first generates category-independent region proposals. The second module is a large convolutional neural network that extracts a fixed-length feature vector from each region. The third module is a set

of class specific linear SVMs.
In spite of its improvements over traditional methods and significant improvements over previous methods their method had a few disadvantages. Firstly, it required each ROI to be computed individually through the CNN as the input size for the CNN was fixed. Secondly, the training process for their architecture is a multi stage pipeline. Lastly, the training is expensive in space and time.

The convolutional layers accept arbitrary input sizes, but they produce outputs of variable sizes. The classifiers (SVM/softmax) or fully-connected layers require fixed-length vectors. To adopt the deep network for images of arbitrary sizes, [8] replace the last pooling layer with a spatial pyramid pooling layer. It reuses feature maps of the last to project region proposals of arbitrary sizes to fixed-length feature vectors.
With spatial pyramid pooling, the input image can be of any sizes. This not only allows arbitrary aspect ratios, but also allows arbitrary scales. The scales play important roles in traditional methods. The coarsest pyramid level has a single bin that covers the entire image and is called global pooling operation.
SPP-Net gives better detection efficiency and gains better results with correct estimation of different region proposals in their corresponding scales. It also reduces the computation required compared to RCNN. But it doesn't overcome the multi stage pipeline and the additional storage required. Additionally, the conv layers preceding the SPP layer cannot be updated with the fine-tuning algorithm.

An accuracy drop of very deep networks is unsurprising in SPP because the conv layers preceding the SPP layer cannot be updated. To this end, [9] introduced a multi-task loss on classification and bounding box regression and proposed a novel CNN architecture named Fast R-CNN.
Similar to SPP-net, the whole image is processed with conv layers to produce feature maps. Then, a fixed-length feature vector is extracted from each region

proposal with a region of interest (RoI) pooling layer. The RoI pooling layer is a special case of the SPP layer, which has only one pyramid level.
All parameters in these procedures (except the generation of region proposals) are optimized via a multi-task loss in an end-to-end way. Fast R-CNN samples mini-batches hierarchically, namely N images sampled randomly at first and then R/N RoIs sampled in each image, R=N where R represents the number of RoIs.
In the Fast R-CNN, regardless of region proposal generation, the training of all network layers can be processed in a single-stage with a multi-task loss. It saves the additional expense on storage space, and improves both accuracy and efficiency.

Region proposal computation is also a bottleneck in improving efficiency. To solve this problem, [10] introduced an additional Region Proposal Network(RPN). It shares full-image conv features with detection network. RPN is achieved with a fully-convolutional network, which has the ability to predict object bounds and scores at each position simultaneously. Similarly, RPN takes an image of arbitrary size to generate a set of rectangular object proposals. RPN operates on a specific conv layer with the preceding layers shared with object detection network.
The network uses a sliding window approach over the conv feature map and fully connects to an n x n spatial window. A low dimensional vector is obtained for each sliding window and fed into two sibling FC layers. It uses anchors of 3 scales and 3 aspect ratios.
It can be fully trained in an end-to-end way. However,RPN produces object-like regions (including backgrounds) instead of object instances and is not skilled in dealing with objects with extreme scales or shapes.

If a candidate box does not correctly overlaps with a true object, the voting score for a particular category should be low.To consider the translation variance, [11] proposes Region-based Fully Convolutional Network.
The R-FCN architecture is designed to classify the RoIs into object categories

and background. In R-FCN, all learnable weight layers are convolutional and are computed on the entire image. The last convolutional layer produces a bank of k*k position-sensitive score maps for each category, and thus has a k*k*(C + 1)-channel output layer with C object categories (+1 for background).
R-FCN ends with a position-sensitive RoI pooling layer. This layer aggregates the outputs of the last convolutional layer and generates scores for each RoI.Each RoI is divided into k x k bins and each of the k k bin aggregates responses from only one score map out of the bank of k k score maps.
The k*k position-sensitive scores then vote on the class of object within the RoI.

Detecting objects in different scales is challenging in particular for small objects.[12] uses pyramid of the same image at different scales to detect objects.
Feature Pyramid Networks(FPN) composes of a bottom-up and a top-down pathway. The bottom-up pathway is the usual convolutional network for feature extraction. As one goes up, the spatial resolution decreases. With more high-level structures detected, the semantic value for each layer increases.FPN provides a top-down pathway to construct higher resolution layers from a semantic rich layer. While the reconstructed layers are semantic strong but the locations of objects are not precise after all the downsampling and upsampling.Lateral connections are added between reconstructed layers and the corresponding feature maps to help the detector to predict the location better.
The bottom-up pathway is composed of many convolution modules each has many convolution layers. As one moves up, the spatial dimension is reduced by 1/2 (i.e. double the stride). The output of each convolution module is labeled as C_i and later used in the top-down pathway.
A 1x1 convolution filter is applied to reduce C_i channel depth to 256-d to create M_i. Then 3x3 convolution is applied to create P_i which becomes the i^{th} feature map layer used for object prediction.
As we go down the top-down path, upsampling of the previous layer by 2 is done,

using nearest neighbors upsampling. Again a 1x1 convolution is applied to the corresponding feature maps in the bottom-up pathway. Then they are added element-wise. A 3x3 convolution again to output the $(i+1)^{th}$ feature map layers for object detection.

To model the corpus of all the captions in the training dataset [13] proposes a generative model: Latent Dirichlet Allocation (LDA) that allows sets of observations to be explained by unobserved groups that explain why some parts of the data are similar. For example, if observations are words collected into documents, it posts that each document is a mixture of a small number of topics and that each word's creation is attributable to one of the document's topics.
Another approach to LDA using Deep Neural Network is given by [14]. It proposes the basic idea of the LDA to DNN knowledge transfer learning by training a DNN model which can simulate the behavior of LDA inference, but with much less computation.

2.2 Outcomes of Literature Review

RNNs were initially proposed for caption generation but they lacked scalability due to their vanishing gradient or exploding gradient problem which worsened as the depth of the architecture increased.

This problem was solved by LSTMs which incorporated memory units that explicitly allowed the network to learn when to forget previous hidden states and when to update hidden states given new information. Also, stacking multiple LSTMs in a model and adding a feedback mechanism between the layers improves recognition and enables end -to-end mapping from input image to sentences.

The Encoder-Decoder framework is implemented by successive CNN and RNN sub models. CNN extracts the visual features from the image and feeds it to RNN

for the sequential generation of sentence describing the image.

Attention Mechanism further enhances the accuracy and robustness of the models by focusing on the salient regions of the image.

LDA allows capture of higher-level semantic information encoded in an image.Being a generative probabilistic model of a corpus, it allows representation of documents as random mixtures over latent topics, where each topic is characterized by a distribution over words.

2.3 Problem Statement

To model an Image Captioning system with region-based attention and scene-specific contexts.

2.4 Objectives and Methodology

- To generate localized candidate regions at multiple scales, which contain visually salient objects, to represent images and generate fixed length feature vectors.
- To extract scene-related global contexts, followed by injecting scene contexts into LSTMs.
- To build an LSTM-based neural network that models the attention dynamics of focusing on those regions as well as generating the words sequentially.
- To integrate individual components and evaluate the results using the BLEU score metric on MS-COCO dataset.

3 Research Methodology

3.1 System Architecture

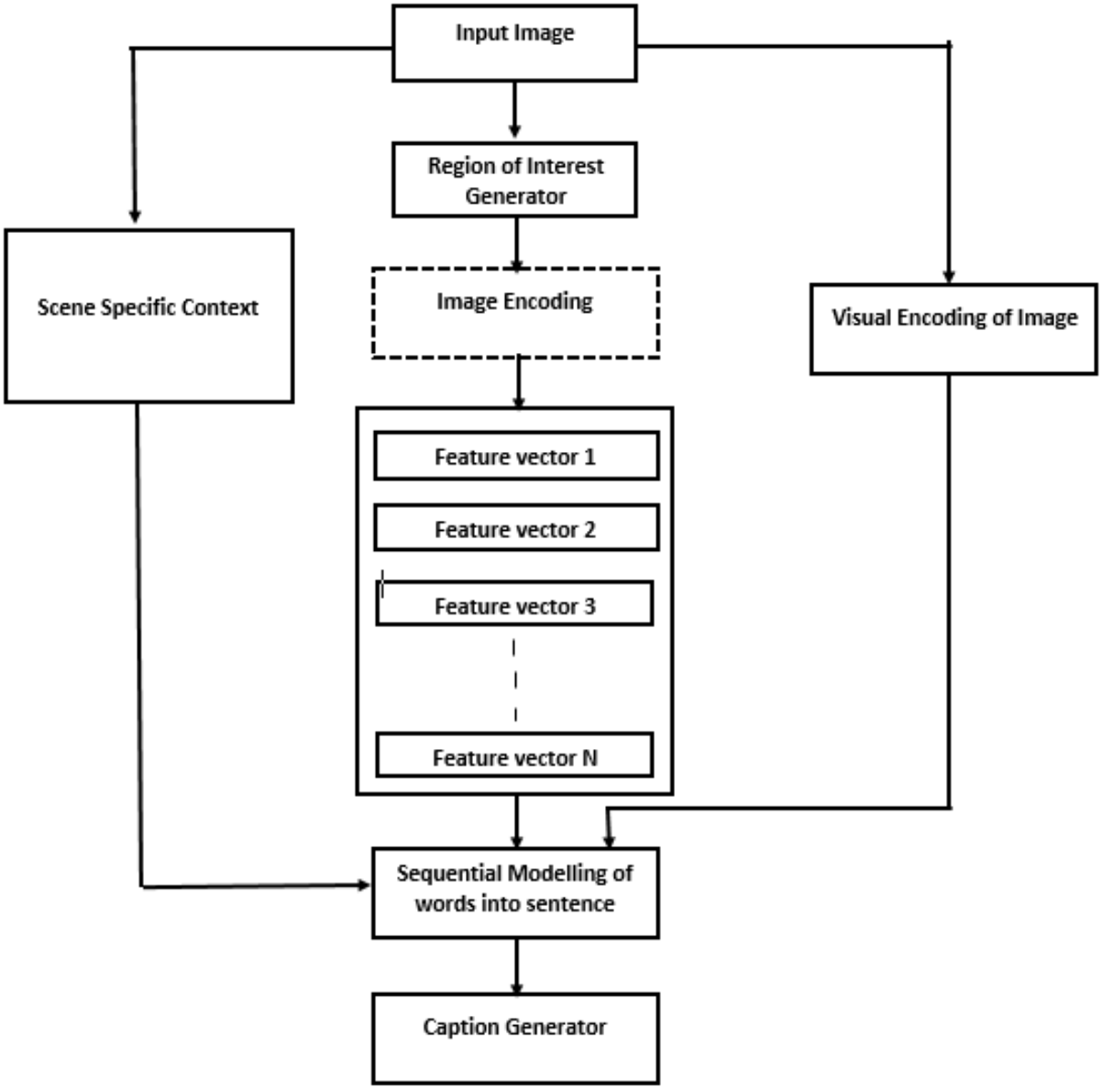


Figure 1: Architecture for Scene Specific Caption Generator with Region based attention

The input is fed into the system and it tries try to capture the information from various parts of the image by generating Regions of Interest(ROIs),scene-specific contexts and a visual encoding of the entire image. This helps us in creating richer encoding of the image, which contains most of the semantic information and hence improves the quality of captions generated. ROIs provide information that is rich both semantically and locally. A feature representation of the entire image aids in capturing the interactions between the ROIs and the entire visual context. Lastly scene-specific contexts contains information about the textual description of the image in form of a vector inferable directly from the image.
All three of these are fed into a sequential model with an attention mechanism. The attention mechanism is used to pay focus on different parts of the image for each word in the output sequence.
The output of this sequential model is the generated caption which translates the information contained in the image to words by exploiting the close correspondence between visual concepts and their textual realization as words in sentences.

3.2 Creating Region of Interests

3.2.1 Selective Search Algorithm

The technique of selective search is used to construct a hierarchical segmentation of the image. The technique first uses color and texture features to over-segment the image, and merges neighboring regions to form a hierarchy of segments until the whole image merge into a single region. For each identified region, a tight bounding box is used to delineate its boundaries.
Among the large number of regions at different levels (i.e.,scales), we select semantically meaningful, primitive and non-compositional but contextually rich ones by training a binary classifier which indicates whether a region is good or bad.

For each image, the top 29 regions are selected according to the outputs of the classifier, and additionally include the whole image to have in total R=30 regions.

Diverse sizes are preferred as not all objects in a image have the same size.

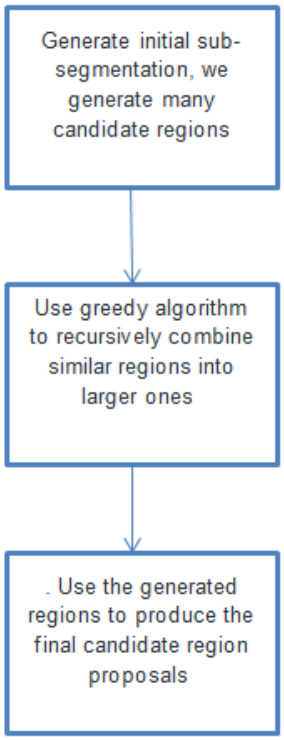


Figure 2: Selective Search Algorithm

Problems with Selective Search:

- It takes a huge amount of time to train the network as you would have to classify 2000 region proposals per image.
- It cannot be implemented real time as it takes around 47 seconds for each test image.
- The selective search algorithm is a fixed algorithm. Therefore, no learning is happening at that stage. This could lead to the generation of bad candidate region proposals.

3.2.2 Faster R-CNN

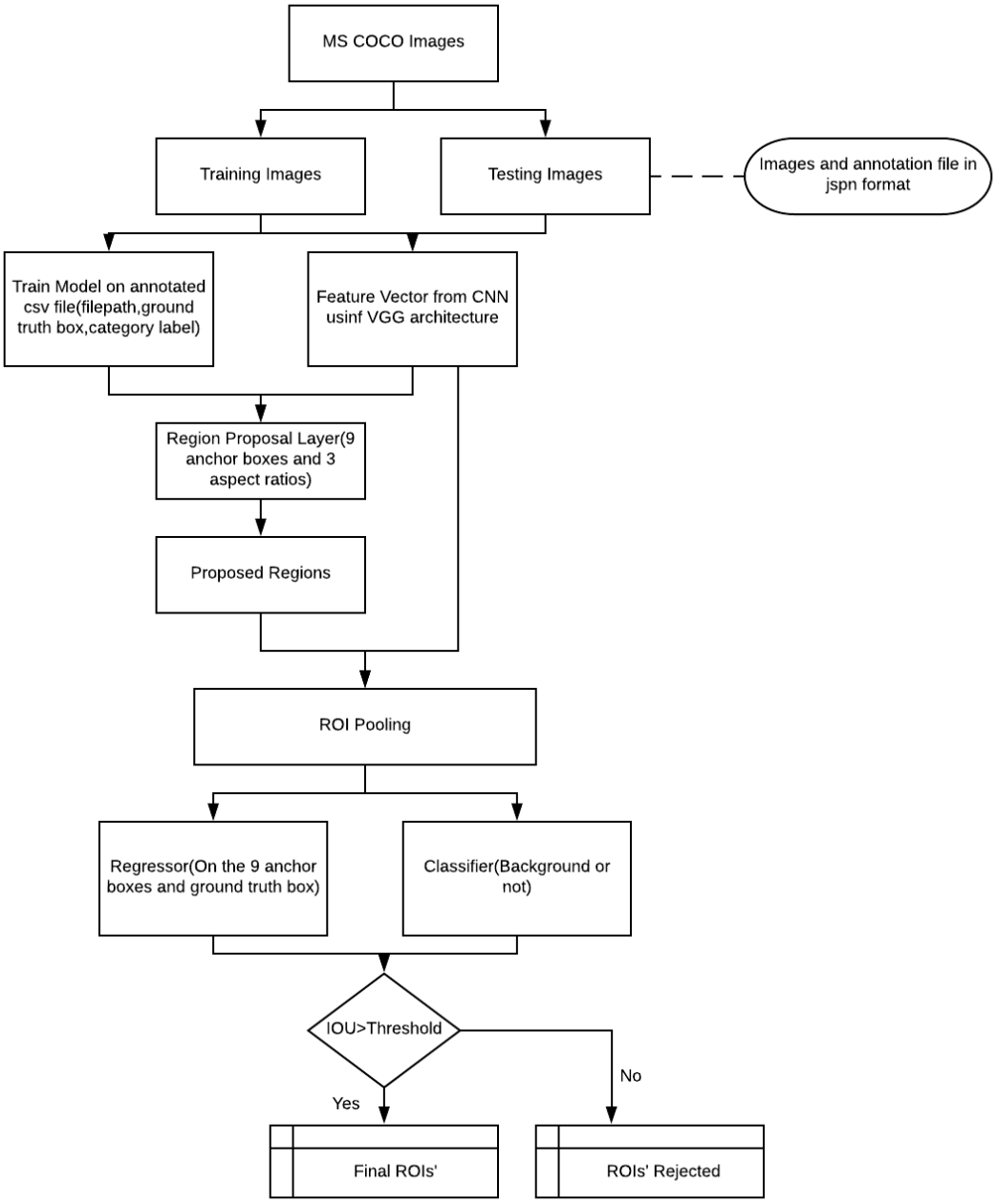


Figure 3: FRCNN work flow diagram

Instead of using selective search algorithm on the feature map to identify the

region proposals, a separate network is used to predict the region proposals. The ground-truth bounding boxes are used as positive examples, and randomly select some patches, which have at most 30 percent intersection with ground-truth, as negative examples.

Each patch is resized into 224x224 to feed into the VGG16 network to obtain 2,048-dimensional CNN features. In our literature survey we found that CNN with higher accuracy on classification task leads to better performance on image captioning task. Thus VGG16 was chosen as the feature extractor.

The model extracts features from objects bounding boxes, which is a more direct representation. The system leverages the intuition that different parts of sentences ought to correspond to different regions on the image. Therefore the system need to model how captioning moves between regions, using an attention model to characterize the dynamics which will be accomplished by the LSTM.

It then trains a CNN having architecture same as VGG16 whose output is fed to the Region Proposal Layer along with the annotate.txt file which contains the coordinates and their corresponding labels. It projects 9 anchor boxes having 3 different aspect ratios to get potential Region of Interests.

These are then fed into a regressor and a classifier which finally gives the Region of Interest. The loss used for classifier is the softmax loss while the mean squared error is used for regressor.

$$softmax(z_{\mathrm{i}}) = \frac{exp(z_{\mathrm{i}})}{\Sigma_{\mathrm{j}} exp(z_{\mathrm{j}})} \tag{1}$$

$$Loss_{\mathrm{i}} = -log(softmax(z_{\mathrm{i}})) \tag{2}$$

$$MSE = \frac{1}{N} \Sigma^{\mathrm{n}}{}_{\mathrm{i}=1} (y_{\mathrm{i}} - y'_{\mathrm{i}})^2 \tag{3}$$

The model has been trained on MSCOCO dataset for 20 epochs. During testing, the ROIs having classifier probability greater than threshold are the required ROIS.

3.3 Decoding using an LSTM and Attention mechanism

Three sets of variables are available: the hidden state $\{h_t\}$ for characterizing the transition of latent meanings governing the transitions from one concept to another, the output variables$\{w_t\}$ for the words being generated, and the input variables $\{v_t\}$ describing the visual context for the image, for example, for the visual element(s) being focused.

For simplicity, the subscript t indices the time, expanding from 0 (START) to T + 1 (END) where T is the length of the sentence. The model is a sequential model, predicting the value of the new state and output variables at time t + 1, based on their values in the past as well as the values of input variables up to time t+1.

At any time t, time system is presented with the image which is already analyzed and represented with R localized patches at multiple scales. At time t, system predicts which visual element is being focused and obtain the right feature vector as visual context by using an one-hidden-layer neural network with R softmax output variables.

The main assumption in sequence modelling networks such as RNNs, LSTMs and GRUs is that the current state holds information for the whole of input seen so far. Hence the final state of a RNN after reading the whole input sequence should contain complete information about that sequence. Attention mechanism relax this assumption and proposes that we should look at the hidden states corresponding to the whole input sequence in order to make any prediction.

$$e_{tj} = a(s_{t-1}, X_j) \tag{4}$$

$$\alpha_{tj} = \frac{exp(e_{tj})}{\Sigma_{k=1}^{N} exp(e_{tk)}} \tag{5}$$

$$c_t = \Sigma_{k=1}^{N} \alpha_{ij} X_j \tag{6}$$

$$s_t = f(s_{t-1}, y_{t-1}, c_t) \tag{7}$$

$$y_t = g(y_{t-1}, s_t) \tag{8}$$

where X_i is the i^{th} input to the decoder, s_i is the output of the i^{th} time step.

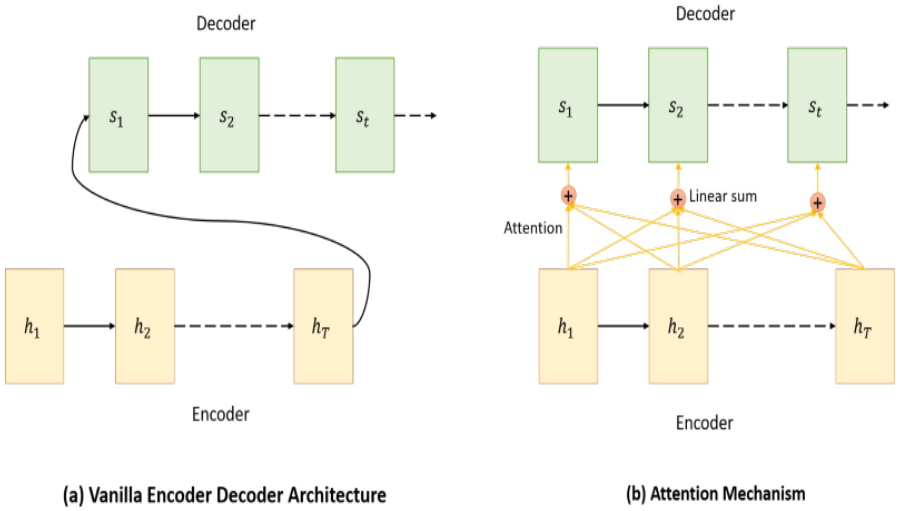


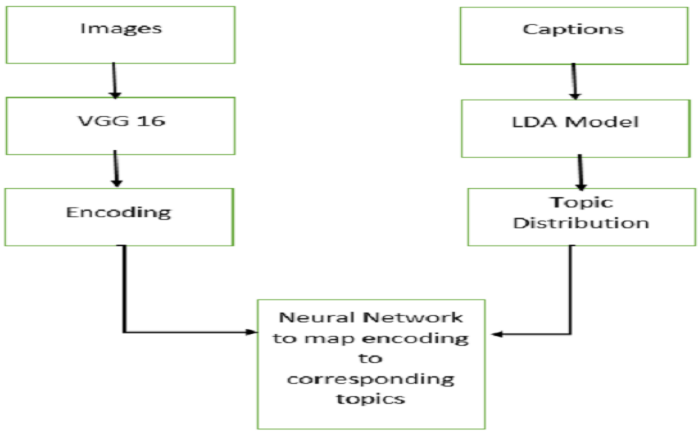
Figure 4: Difference between attention and normal encoder decoder architecture

3.4 Injecting Scene Vectors

The ideas used will be : 1) Unsupervised clustering of captions into scene categories and 2) Supervised learning of a regressor to predict the scene vectors from the visual appearance.

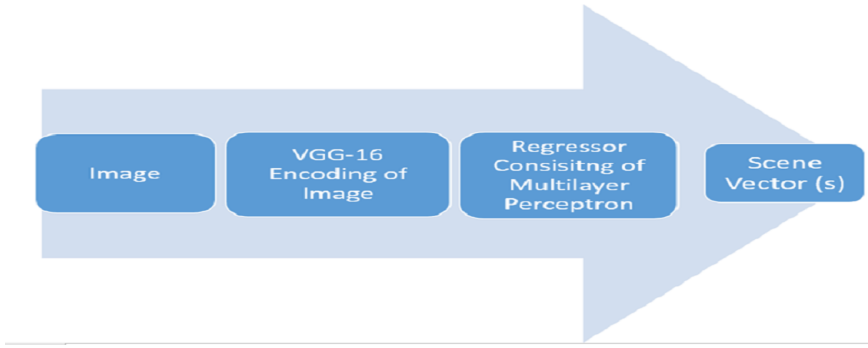
Unsupervised Clustering : For each image, a multi-dimensional topic vector is obtained that assigns its caption into the memberships of multiple categories called scene vectors. Scene vectors are inferred from the corpus that is generated using the captions of training dataset by using LDA.

Supervised Learning : The clustering is followed by the training of a regressor for prediction of scene vector when presented with an image. The training samples for this regressor are the images from the training dataset with the target outputs being the inferred scene vectors. A multilayer perceptron will be used as a regressor. In case when the captions are not available, global feature vectors of the images will be used.



Encoding results in 4096x1 vector which is fed alongwith 20x1 vector of topical distribution as output for training to the neural network.20 here is the number of topics

Figure 5: LDA Flow Diagram - Training Phase



The regressor shown in Figure 5 consists of the multilayered perceptron trained on the scene vectors obtained from captions.txt of the MSCOCO dataset using LDA.

Figure 6: LDA Flow Diagram

LDA model is to supervise the training of a deep neural network (DNN).The number of output units are equal to the number of topics and the number of hidden layers are three and two times of the output units for the first and second hidden layer, respectively in the DNN. The hyperbolic function is used as the activation function.

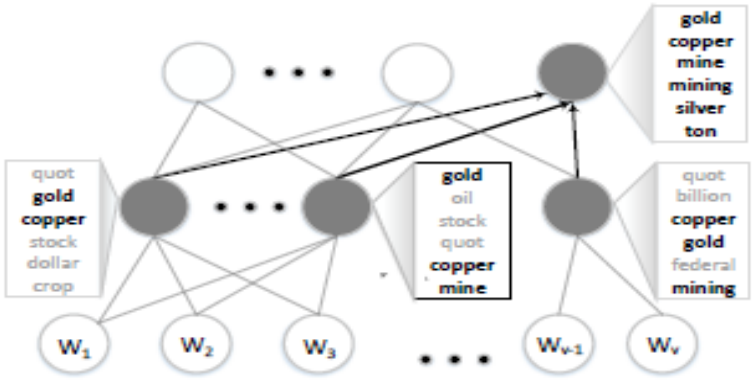


Figure 7: LDA using the DNN architecture

The scene vectors will be then injected into LSTM and the sentence generation process will adapt to be scene-specific by biasing the gates in the LSTM.

3.5 Benchmarking and Analysis

The image captioning system is evaluated on MSCOCO data set.The Microsoft Common Objects in Context (MS COCO) dataset contains 91 common object categories with 82 of them having more than 5,000 labeled instances. In total, the dataset has 12,000 labeled instances out of which 8,000 were used for training and rest for validation.MS COCO has fewer categories but more instances per category. A quantitative analysis of the results is done by evaluating the obtained results using the cumulative BLEU (bilingual evaluation understudy) score. BLEU score is a metric for evaluating a generated sentence to a reference sentence. A perfect match results in a score of 1.0, whereas a perfect mismatch results in a score of 0.0.

Scores are calculated for individual translated segmentsgenerally sentencesby comparing them with a set of good quality reference translations. Intelligibility or grammatical correctness are not taken into account. This value indicates how similar the candidate text is to the reference texts. Few human translations will attain a score of 1, since this would indicate that the candidate is identical to one of the reference translations. For this reason, it is not necessary to attain a score of 1. Because there are more opportunities to match, adding additional reference translations will increase the BLEU score.

4 Results and Discussion

- The selective search technique was studied and results were compared by varying the 'k' value which governs the size of bounding boxes which in turn determine the size of the objects that the program is looking for in the image. An optimum balance was difficult to reach here as there was a trade-off between object size and number of bounding boxes. Hence other techniques were researched further.

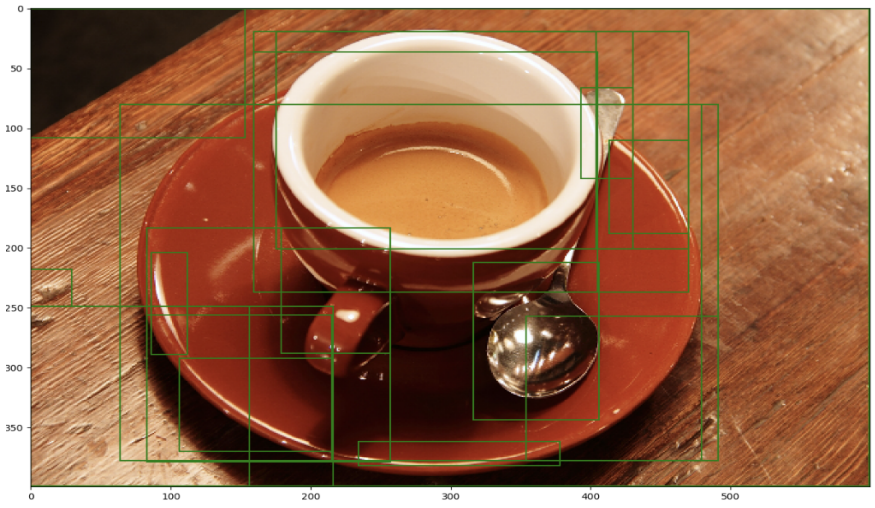


Figure 8: Bounding Boxes for k = 1000

- Latent Dirichlet Allocation has been implemented using the "scikit learn" inbuilt library.The probabilistic measures to redistribute the words among the topics utilize the use of term-frequency(tf) and inverse-document frequency(idf).The model has been trained on the Flicker8K captions.
  The tf-idf measure is used so that the distribution is evenly weighed by all the input words rather than only the frequently occurring ones. Also, as the number of topics are increased for this dataset, the distribution of words becomes more sparse and hence the probability distribution of the words on the whole is suppressed.

- The accuracy obtained with LDA using Deep Neural Networks comes out to be 59

Topic Number	Word 1	Word 2	Word 3	Word 4	Word 5
1	dog	snow	brown	running	pink
2	playing	girl	croquet	little	ball
3	high	doing	air	ski	jump
4	man	chap	wearing	leather	camera
5	dog	walk	large	sand	brown
6	boy	smile	little	shirt	camera
7	red	shirt	near	man	bandana
8	ball	dancing	holding	young	green
9	red	riding	dog	helmet	wearing
10	running	lying	green	men	looking

Table 1: LDA Topic Distribution for 10 topics

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
0.81998967	0.0200122	0.02000074	0.02000056	0.02000246
Topic 6	Topic 7	Topic 8	Topic 9	Topic 10
0.02000009	0.02000065	0.0200058	0.0200024	0.02000083

Table 2: Topic distribution for the sentence: The dog jumping the fence.

In order to further optimise the runtime of the model, LDA was also implemented using the "gensim" inbuilt library and also using deep neural networks. The measures used were the same tf-idf measure.
A deep neural network was implemented that mapped encoded images to a topical distribution vector as output for the scene specific context. The deep neural network had 2 hidden layers with number of nodes in the output layer equal to number of topics. The number of topics was kept 20 for now as mentioned in the literature.

- VGG16 model has been implemented using the models from Keras and a fixed length vector of size 4096 is generated by removing the last classification layer of the pre-trained model. This feature vector will be fed into the LSTM along with the Regions of Interests to provide global context to a sequencing model.

- FRCNN model has been implemented using Keras. During training, it takes the images, the bounding box coordinates and the corresponding ground truth labels of objects in the image as the input.
  The model has been trained on MSCOCO dataset for 15 epochs having epoch length of 50. Adam optimizer has been used for both the classifier and regressor. The loss used for regressor is Mean Absolute Error(MAE).

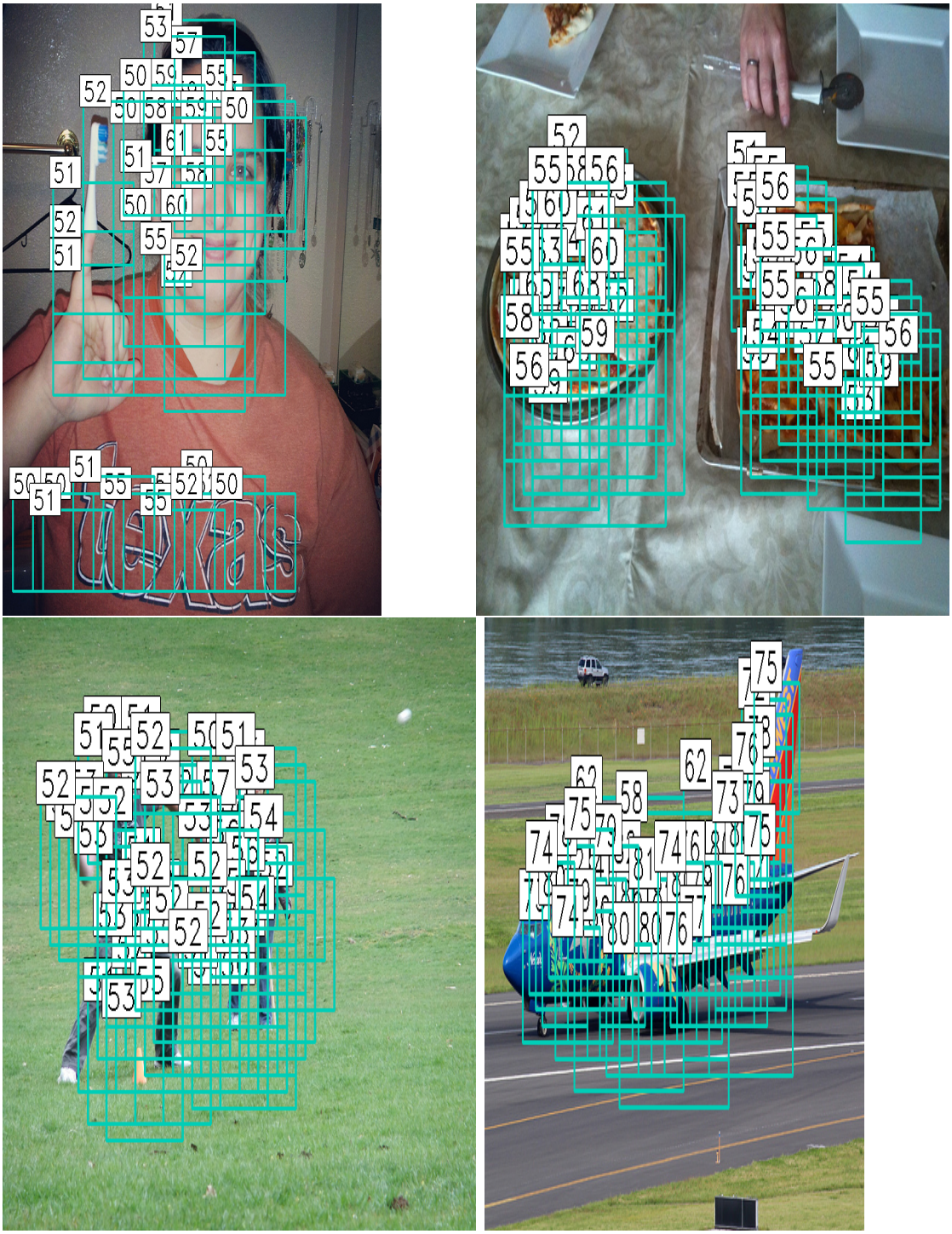


Figure 9: Output Images from Faster RCNN's RPN

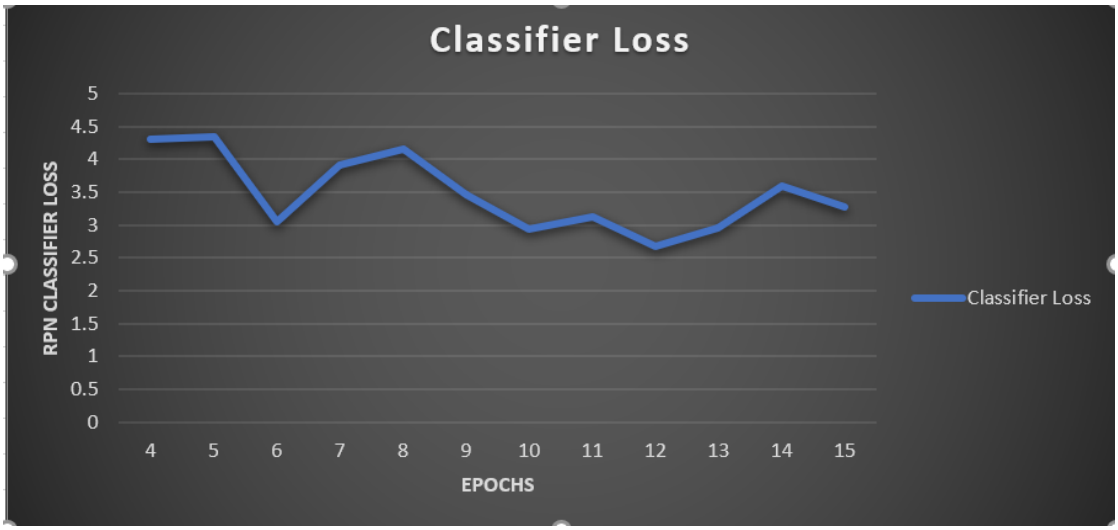


Figure 10: RPN Classifier Loss for lr = 1e-5

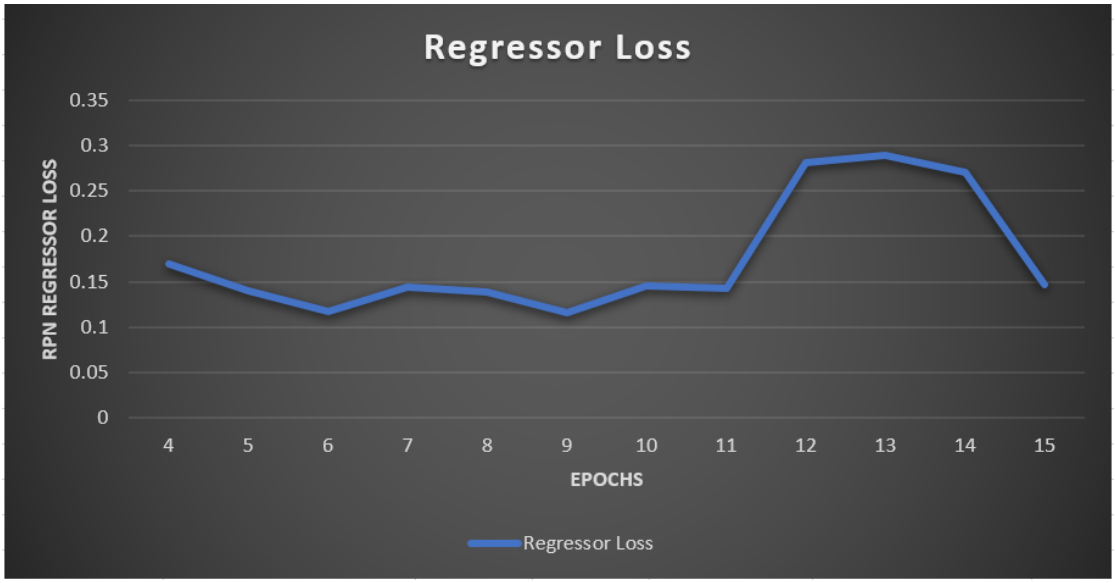


Figure 11: RPN Regressor Loss for lr = 1e-5

Reference Caption : A living room filled with furniture and a rug.

Predicted Caption : A living room with a rug and table.

BLEU Score : 0.2678943015

Reference Caption : A bus driving in a city area with traffic signs.

Predicted Caption : A city bus driving down a street next to a tall building.

BLEU Score : 0.226329821469

Reference Caption : A man riding skis down a snow covered path.

Predicted Caption : A man flying through the air riding skis.

BLEU Score : 0.213048496454

Reference Caption : A small loaded pizza on a yellow plate.

Predicted Caption : A pizza with cheese and tomatoes on it.

BLEU Score : 0.2691677134612

Figure 12: Input Image, Referenced and Candidate caption, Bleu score

5 Conclusion and Future Work

The project aim to propose an image captioning system that exploits the parallel structures between images and sentences. The model aims to align the process of caption generation and attention shifting among the visual regions. It also is able to introduces the scene-specific contexts to LSTM that adapts language models for word generation to specific scene types.

For future work, we can try various architectures of RNN to improve the performance of the model either in terms of accuracy or in terms of memory/space usage. LDA can be further trained on different number of topics to understand the comprehensiveness of the model as per the dataset.

By reducing the memory usage this model can be easily deployed on mobile devices. The Quality of Service of applications having a visually impaired audience can be improved by using this model by reducing the time overhead and giving real-time descriptions of their surroundings to their users.

By improving the accuracy and extracting information from the attention mechanism we can have better localization of the objects in the caption. This will be particularly useful tasks involving image annotation.

Lastly, we can make the model more flexible and versatile by training it not only to predict captions in English but for different languages around the world.

6 Timeline of the Project

SEVENTH SEMESTER

August — Researching and finalizing the problem statement.

Submission of the first proposal

September — Getting started with the literature review and analyzing many papers from computer vision Journals and IEEE

October — Designing a methodology based on our literature review

November — Generation of a rough architecture about the overall flow of the proposed solution and Implementation of selective search and global context vector

EIGHTH SEMESTER

January — Implementation of DNN for LDA modelling and training of FRCNN on MSCOCO

February — Implementation of Attention mechanism in LSTM

March — Final integration of individual components using Attention mechanism in LSTM

References

[1] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, Show and tell: A neural image caption generator, in Proc. IEEE Conf. Comput. Vis.Pattern Recognit., 2015, pp. 31563164.

[2] J. Donahue, et al., Long-term recurrent convolutional networks for visual recognition and description, in Proc. IEEE Comput. Vis. Pattern Recognit., 2014, pp. 26252634.

[3] J. Mao, W. Xu, Y. Yang, J. Wang, and A. L. Yuille, Explain images with multimodal recurrent neural networks, in Proc. Int. Conf.Learn. Representations, 2015.

[4] K. Xu, et al., Show, attend and tell: Neural image caption generation with visual attention, in Proc. 32nd Int. Conf. Mach. Learn., 2015, pp. 20482057.

[5] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo, Image captioning with semantic attention, in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2016, pp. 46514659.

[6] J. R. Uijlings, K. E. van de Sande, T. Gevers, and A. W. Smeulders,Selective search for object recognition, Int. J.Comput. Vis.,vol. 104, no. 2, pp. 154171, 2013.

[7] Ross Girshick Jeff Donahue Trevor Darrell Jitendra Malik "Rich feature hierarchies for accurate object detection and semantic segmentation"

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun, "Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition"

[9] Ross Girshick, "Fast R-CNN"

[10] Shaoqing Ren, Kaiming He, Ross Girshick, Jian Sun "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks"

[11] Jifeng Dai, Yi Li, Kaiming He, Jian Sun "R-FCN: Object Detection via Region-based Fully Convolutional Networks"

[12] Tsung-Yi Lin, Piotr Dollar, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie "Feature Pyramid Networks for Object Detection"

[13] David M. Blei, Andrew Y. Ng, Michael I. Jordan "Latent Dirichlet Allocation"

[14] Dongxu Zhang, Tianyi Luo, Dong Wang "Learning from LDA Using Deep Neural Networks" in International Conference on Computer Processing of Oriental Languages