DEEP LEARNING BASED IMAGE CAPTIONING OVER MULTIMEDIA IMAGES

Dipan M. Arya
AI, Data Science & Machine Learning
Accenture
Ahmedabad, India

*ABSTRACT - Image classification is the process of correctly identifying and grouping of a set of images correctly. This requires the capturing of significant elements, their innate qualities, its relationships in a particular image. The captions generated for a*n identified is called image semantics and is process of generating an apt caption for any interpreted image. The primary focus of research is on allowing computers to interpret and recognize pictures in much the same way humans do. It also presents additional challenges that are both time-consuming and expensive. Semantic correctness and syntactic veracity are the key attributes to gauge image classification models. Researchers are currently coming up with better approaches using computer vision to match the requirement and produce semantically correct descriptions after identifying images correctly. This study presents a comparison of image classification models for image captioning using deep learning techniques maximizing the strength, overcoming limitations and presents an analysis of its performance. The evaluation metrics captures the computational accuracy, analyzing the predictive capacities of each deep learning model. A cross section of the MS Coco data set was analyzed using evaluation metrics widely used for deep learning based automated image semantics. The strengths performances and limitations of each model are compared based on the individual model accuracy scores.

Keywords— BLEU, CBIR, CIDEr, METEOR, SC-NLM, VGGNet, RNN, Inception-V3, Resnet-152

INTRODUCTION

1.1 *Background of the study*

Humans encounter a large number of images daily from multiple sources. A simple human routine of car driving would expose drivers and passengers to large billboards with advertisements. Texting or answering calls on their cell phones will show users their profile picture. (Zakir Hossain et al., 2018) Human minds easily analyse the images with seamless integration with the visual stimulus (eyes), semantic understanding (textual or speech) and relative contextualisation in the brain in perfect symphony. People can quickly provide a caption for an image understanding and relating it to the visual stimulus shown. Identification of the object or image, describing the image and placing it in context based on past knowledge, connecting with the present circumstance and verbally read out a near-perfect description for the picture shown. Replication of this marvellous human ability is an extremely challenging aspect if the same is expected from machines. Machines having this ability would be a tremendous achievement and will have several applications. Amazing transformations can take place in the human realm. It could help the visually impaired quickly identify and understand the visual world and make good to extant their inability to see and comprehend the visual world. A simple act of visual captioning transferred to machines could herald the way for advanced robotics with speech recognition, visual understanding, and automatic responses, much like a human conversation, and make robots more mobile and more integrated into the visual world. (Cascianelli et al., 2018) This amazing capability integrated into machines can see its use in the automotive sectors, defense industry (Shi and Zou, 2017), healthcare, e-commerce, social media (M. Banko, V. Mittal, 2000) heralding its adoption in more fields opening the doors for limitless possibilities.

Image captioning is the amalgamation of Deep Computer Vision (CV), and Natural Language Processing (NLP). Understanding the visual context and providing an accurate description of the image are the two main components: enable machines to possess this ability. (Khurana and Mundada, 2018) Researchers developed CNN-RNN frameworks for image captioning. However, they soon realised this methods drawback: each image caption would have equal importance with individually defined importance when one image should have a higher weightage over the others. The other major drawback is an inaccurate description of image captions. CNN-RNN's is an Encoder-Decoder system with CNN being a Convolutional Neural Network that uses this deep learning technique for feature extraction and object detection. The decoder generates words for the images and when both the techniques are used in tandem becomes an image captioning system for visual comprehension and textual description.

Retrieval and template-based methods are other approaches used by researchers, for the retrieval-based approach as the name suggests, it's based on retrieval of existing captions, here the system closely observes and identifies a query image from a training dataset which then picks a caption from the available pool of captions. Retrieval systems work in the following steps: (J. Curran, S. Clark, n.d.; Lin, n.d.; v. Ordonez, G. Kulkarni, 2011)

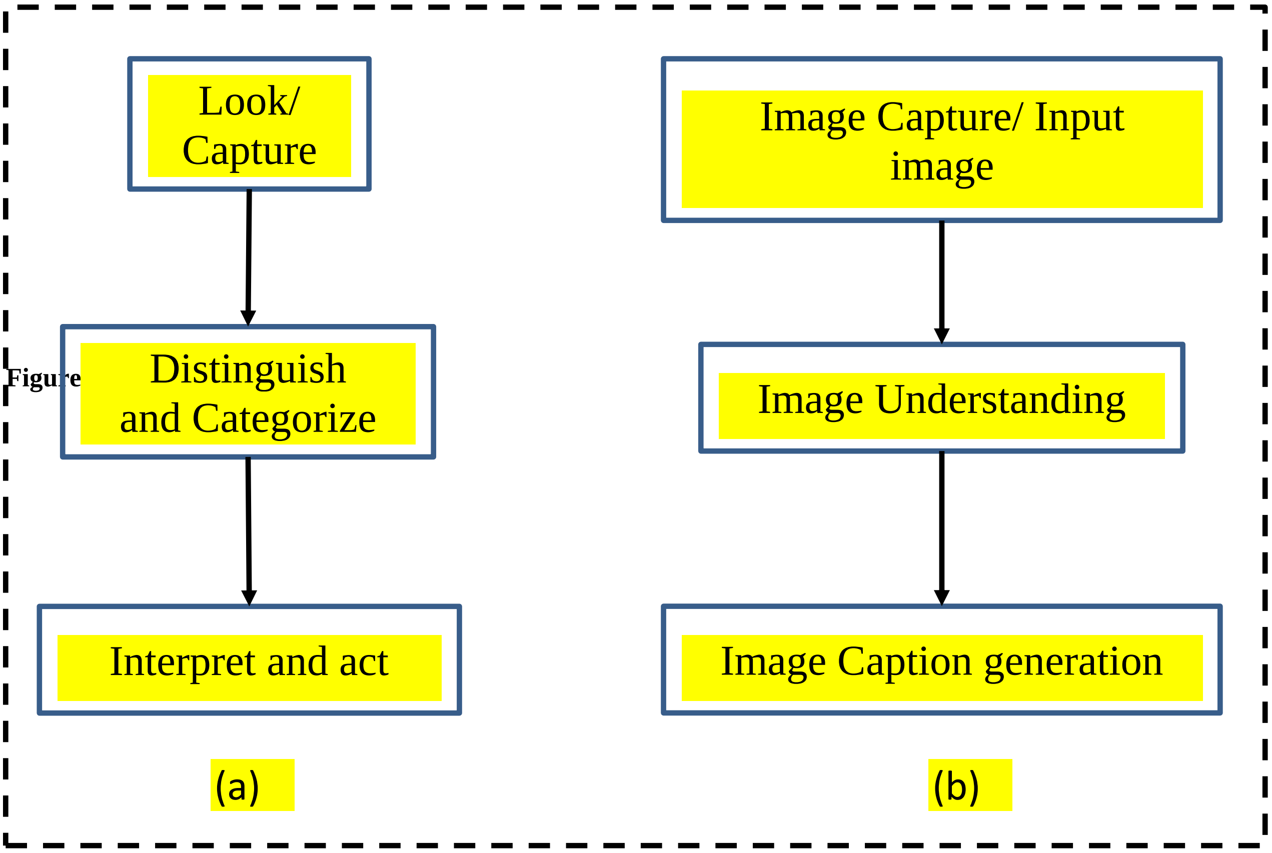
1. Every query image is plotted into the meaning space by solving a Markov Random Field
2. Semantic distance between existing images in the data set is deduced by Lin similarity measure
3. Existing sentences are parsed with the help of Curran and Clark Parser
4. The image caption will be selected, which is the nearest caption to the queried image

Image Captioning using set templates with a number of blank slots is called a Template-Based image captioning system. Blank slots are filled with the detected objects, attributes, and behavior from an image. This process works using a triplet where the image object, action and scene attributes are taken to fill the blank slots. Both the methods have their limitations since they are restricted to the limited set of templates and captions in retrieval systems which is a serious drawback unable to have the flexibility to observe and caption newer mixtures of images. Providing grammatically correct captions without newer combinations of images has led researchers to explore newer multimodal methods. The multimodal space technique mainly follows the below steps:

1. Advanced Deep Learning models and multimodal language models both learn image and textual contents jointly and try them in a multidimensional space.
2. Immediately next step generates semantics based on provided details in the First Step LSTM a Recurrent Neural Network (RNN) is utilised for deep learning as a decoder due to its advantages – long term memorising capabilities. Comparing existing algorithms based on its predictive capabilities for image captioning is the purpose of this study.

1.2 *Problem statement of the study*

The main area of research is in enabling machines in understanding and identifying images just like a human looks at images and comprehends them. This also brings in additional challenges which are time-consuming and expensive. The caption tries to capture the understanding of the image but not the language. Currently, researchers are coming up with better approaches using computer vision that can match the requirement and are able to produce semantically correct descriptions after identifying images correctly. The major problems that arise in image captioning can be summarized into three types. (Overcoming Challenges In Automated Image Captioning, 2021)



Figure

Comparison of image captioning with (a) representing humans and (b) representing machines.

1.2.1 *Compositional nature of visual scenes*

The first obstacle is the compositional nature of natural language and visual scenes. Although the training dataset comprises occurrences of such objects in their contexts, a captioning system should be able to generalize by writing objects in various settings. Traditional captioning systems lack compositionality and naturalness since they often construct captions in a predefined sequence, with the next generated word reliant on the sequential order of the previous word along with the image feature. This frequently leads to syntactically correct yet ssemantically meaningless language structures, as well as a lack of diversity in the captions generated by such systems.

1.2.2 *Generalization problem due to data set bias*

Dataset bias is the second problem that existing captioning systems face. Overfitting to familiar entities that co-occur in a common context (e.g., bed and bedroom) causes systems in failing to generalize the scenes where the same objects appear in unknown contexts (e.g., bed and forest). Ensemble models can be investigated to minimise bias, but they cannot always be relied on due to data complexity idiosyncrasies.

1.2.3 *Performance evaluation*

The quality of produced captions is the third challenge. Using automated metrics, although useful in some cases, is still unsatisfactory because they ignore the picture. In many cases, their scoring is insufficient and even misleading, especially when it comes to scoring diverse and descriptive captions. In terms of scoring captioning systems, human assessment remains the gold standard.

The study of image captioning has taken a long time and has gone through several stages based on various technologies. The application of neural network technology to image captioning research has created a new situation in recent years. Although the neural network's powerful data processing capability has produced outstanding results in the study of image captioning generation, there are still some issues to be resolved. Some of the challenges arising out of the complexity of images for example a model not being able to identify the number of objects but did identify the object accurately. The present capability of models is capable to describe objects in more detail that are simple and not complex. Challenges arise when numerous objects are present and complex object relationships exist, models are prone to errors and inaccurate description. At times the models pay more attention to unimportant details and less importance to important image aspects thereby giving a fallacious description.

Training datasets heavily impact the image captioning capability of models leading to overfitting and bias creeping into the description output of the models. The template-based and retrieval systems are reliant on the quality and diversity of the training data sets. The network is fed a real word vector or a mixture of real words and images at each time phase during the training process. The expected word is the network's predicted word. In the test phase, however, the network inputs at each time step are the output word vectors from the training dataset's vocabulary. When an image includes novel objects, the nearest object in the data set is chosen rather than the true object. As a result, when new objects are produced, there are inconsistencies in the training and testing process. As a result of these inconsistencies, cumulative error sampling may be generated, and text explanations that are entirely inconsistent with the image content may be generated, resulting in incorrect description results. Hence there will be a mismatch in training and test scenarios, which needs to be addressed by better techniques that can reduce this gap.

The latest deep learning or machine learning-based image captioning system necessitates a large number of marked training samples. In practical applications, it is essential to include a text explanation for the picture in a variety of languages to satisfy the needs of different native language users. Today, many training examples are listed in English and Chinese documents, but there are few mark-ups of explanations in other languages. Manual labelling would take a lot of manpower and time if the textual explanation of each language in the picture is finished. Cross-language text captioning is a very big problem area for further research.

1.3 *Aims and objectives*

The current work aims to compare the state-of-the-art algorithms in image classification for captioning and analyze the results based on the chosen evaluation metrics for such studies. Each model strengths, weaknesses and limitations will be investigated and presented for further analysis. This study should help in choosing the right transfer learning model for large datasets for its inherent strengths based on model architecture and predictive capacities.

Below is the list of objectives that would be achieved as part of this research:

1. To select the right type of classification models for image captioning
2. To compare each model efficiency for accuracy
3. To judge model accuracy using evaluation metrics popularly used for classification and captioning studies

1.4 *Scope and significance of the study*

Caption generation is a fascinating artificial intelligence problem that involves generating a descriptive sentence for a given image. It uses two computer vision techniques to understand the image's content, as well as a language model from the field of natural language processing to convert the image's understanding into words in the correct order. Image captioning has a variety of uses, including recommendations in editing software, use of virtual assistants, image indexing, accessibility for visually disabled people, social media, and a variety of other natural language processing applications.(Srinivasan and Sreekanthan, 2018) The effects of deep learning approaches are cutting-edge. Deep learning models have been shown to be capable of achieving optimal results in the field of caption generation problems. A single end-to-end model can be specified to predict a caption provided a picture, rather than requiring complex data preparation or a pipeline of explicitly built models. (Biswas et al., n.d.)

1.5 *Structure of the study*

The goal of the research is to improve the accuracy of image captioning systems by employing innovative approaches and algorithms in picture detection and captioned. The subject is carefully investigated, and we look at all of the novels approaches that researcher who are knowledgeable with this field have applied to overcome these issues. Recent papers that are published on new algorithmic development that are able to address problem areas by using a newly developed model to see improvements are also considered. We also will study papers where a combination of methods is used in tandem to improve the overall efficacy of generated captions. To improve the overall capabilities of the image detection and captioning systems, we will select the appropriate and are useful based on time, effort, and accessibility on the MS Coco data set. In the next chapters, the results will be presented in order to validate and pick the best approaches based on performance in order to choose the most efficient image classification model for image captioning studies.

2.1 *Evolution and Historical Context*

There will be a thorough and in-depth review of chapters to understand various new developments taking place and collating various contributions from various researchers. We describe and depict the main classes of current picture subtitling strategies in this section, which include layout-based picture captioning CBIR-based Image Captioning, and novel semantics generation techniques. Layout-based methodologies have skeleton templates with various DOT Spaces to form a semantic. In these methodologies, various items, activities, attributes are distinguished at the initial level and afterwards, the DOT Spaces in the layout are occupied. Yet, the Skeleton of layouts are predefined and can produce fixed-length semantics. Apart from this, parsing models are also used which covers syntax, semantics and pragmatics in terms of probabilities among parses that have been introduced which are more impressive than the fixed skeleton layout-based method (Farhadi et al., n.d.). The other approach is a retrieval-based technique in which Captions are retrieved and constructed using a set of existing available semantics. The content retrieval-based procedures use existing images and discover similar images based on the extracted features. Along with shortlisted images, it also uses available associated semantics to images and generates semantics of input images (Gong et al., n.d.; He et al., n.d.; Ordonez et al., n.d.; Sun et al., n.d.).

Apart from these, researchers are also using the multimodal space (Kiros et al., n.d.) to learn a dense feature embedding against each word. In the multimodal space technique, the extracted image features are mapped to the word features into a common space.

The multimodal space technique mainly follows the below steps:

1. Advanced Deep Learning models and multimodal language models, both learn image and textual contents jointly and try it in a multidimensional space.

2. Immediately next step generates semantics based on provided details in the First Step.

Following the multimodal space as the groundwork, the research has also extended and use the LSTM for sentence encoding. This leads to generating a realistic image caption over the approach presented earlier. The new combination of components is known as the Structure Content Neural Language Model (SC-NLM) which can generate semantics more accurately (Kiros et al., 2014).

In addition to this, researchers also used the RNN with multimodal space which is known as the m-RNN model to generate the semantics of an image. In this technique, the Images are provided to Convolutional Neural Network as an input and image descriptions are provided to the Recurrent Neural Network. These two neural networks are inside collaborating with one another and create the likelihood of the following word to frame a semantic of an input picture completely. This technique uses the strength of the Recurrent Neural Network more effectively. Thus, producing results in a more precise and proficient way to deliver Semantics (Mao et al., n.d.).

Researchers are also focusing on Supervised Learning-Based Image Captioning.

Supervised Learning-Based Image Captioning demonstrated state-of-the-art results on caption generation problems. SL-Based Image Captioning is the machine learning technique to learn and fine-tune a function that maps semantics to an extracted image feature based on training image and semantics pairs. It infers a function from labelled training data consisting of a set of training examples with the use of Encoder-Decoder Architecture, Attention-Based Mechanism. Generally, CNN (Convolution Neural Network) and RNN (Recurrent Neural Network) signifies Encoder and Decoder respectively. The functionality of Encoder is to extract significant features from an input image and the use of Decoder is to generate words – and based on provided extracted image features during an image-encoder phase, generates the grammatically correct sentence that describes the image very well. The first most popular choice for the encoder to extract features from an image is VGGNET, preferred for the simplicity of the model for its power (Ren et al., 2019). The VGGNet is popular and preferably used among the researchers. The other most demanding Encoder is RestNet due to its computational power which makes it unique and more effective compare to all other Convolutional Neural Networks as per past research papers outcomes. (Ren and Hua, 2018) The LSTM is a Recurrent Neural Network (RNN) architecture that is frequently employed as a Decoder in the field of advanced deep learning, with the objective of processing whole sequences of input semantics at once. Due to its long-term memorizing capabilities, LSTM is currently considered as the most well-known technique for image semantics because of its effectiveness in remembering long-short term dependencies by using memory cells. Without a doubt this requires huge storage space and follows a very composite process to create and maintain(Ren and Hua, 2018; Ren et al., 2019)

The other famous approach among researchers is the use of attention mechanisms for image semantics. The attention mechanism is an input processing technique for neural networks that allows the network to focus on specific aspects of an input image and post create a good description that can explain the relationships among the identified objects considering the region of space. There are a couple of ways by which analysts have attempted to copy it, which are broadly known as hard or soft attention mechanisms. The researchers also suggest a novel technique using an attention mechanism which is more focusing on the region space among identified objects along with semantics. The combinations of pairs for a single input image are

-identified object sequences “q”, the sequence of words in associated semantics of an image “w” and at last, set of region space vectors among identified objects in image “v”. In this method, the image region of objects from which object tags are associated and has higher importance than the other regions in an image and based on that caption of an image is generated (Li et al., n.d.).

2.2 *Research papers review and advancements*

(Wang et al., 2020) have discussed the widely used methods for image caption generation which are mainly feature extraction methods and neural network methods. Feature extraction methods use CNN – Convolutional Neural Network which turns the image caption generation into an optimization problem and recognises the image. RNN – Recurrent neural network and LSTM is a type of RNN which is preferred for natural processing tasks. The encoder-decoder approach has gained a lot of credence with CNN fitting the role of encoder and RNN (using LSTM) the role of the decoder. They have given a good comparison of attention mechanisms that mimics human attention spans and focus on the primary information ignoring secondary information.

Attention Name	**Method**	**Comment**
Soft attention	Give a probability according to the context vector for any word in the input sentence when seeking attention probability distribution	Parameterization Derivative enable Definitel
Hard attention	Focus only on a randomly chosen location using Monte Carlo sampling to estimate the gradient	Randomly On the base of probability Simple
Multihead attention	Linearly projecting multiple pieces of information selected from the input in parallel using multiple keys, values, and queries	Linear projection Parallel Focus on information from different representation subspaces in different locations Multiple attention head
Scaled dot- product attention	Execute a single attention function using keys, values, and query matrices Considering	High speed Save space
Global attention	Considering the hidden layer state of all encoders, the weight distribution of attention is obtained by comparing the current decoder hidden layer state with the state of each encoder hidden layer	Comprehensive Time- consuming Large amount of calculation
Local attention	First find a location for it, then calculate the attention weight in the left and right windows of its location, and finally weight the context vector	Reduce the cost of calculations Solve
Adaptive attention	Define a new adaptive context vector which is modelled as a mixture of the spatially attended image features and the visual sentinel vector. This trades off how much new information the network is considering from the image with what it already knows in the decoder memory	Solve when and where to add attention in order to extract meaningful information for sequence words

Semantic attention	Select semantic concepts and incorporate them into the hidden state and output of the LSTM	Optional Merge From top to bottom from bottom to top
Spatial and channel-wise attention	Select semantic attributes based on the needs of the sentence context	Multiple semantics In order to overcome the problem of overrange when using the general attention
Areas of attention	Modelling the dependencies between image regions, title words, and the state of the RNN language model	Interaction Comprehensive

Table 2.1 Table showing attention methods

(Waghmare and Shinde, 2020) attempted to comprehend a hybrid method explaining the use of a Convolutional Neural Network (CNN) to produce accurate image descriptions and the use of an LSTM to correctly assemble descriptive sentences using keywords that had been omitted or extracted. The CNN compares the target image to a vast dataset of training images and attempts to produce an accurate summary using the captions that have been learned.

(Feng et al., n.d.) have crawled a large-scale image summary corpus of two million natural sentences. Their experiments have demonstrated that their proposed model will yield positive results without the use of caption annotations. They presented three training objectives: 1) produced captions are indistinguishable from sentences in the corpus, 2) the image captioning model conveys the object information in the image, and 3) image and sentence features are matched in the common latent space and perform bi-directional reconstructions from each other. Shutterstock was also used to compile a large-scale picture definition corpus of over two million sentences to aid the unsupervised image captioning process. The results of the experiments show that the proposed approach will yield promising results without the use of labelled image-sentence pairs.

(Xu et al., 2015) have described how to train a model both deterministically and stochastically by optimising a variational lower bound. They also demonstrated how the model can automatically learn to focus its eyes on important objects while producing the correct words in the output sequence by visualisation. On three benchmark datasets: Flickr8k, Flickr30k, and MS COCO, they validated the use of attention with state-of-the-art results. Attention based methods gave the best performance on all three bench mark datasets. It shows how learned attention can be used to improve the interpretability of the model generation process, as well as how the learned alignments match human instincts very well. Attention based encoder- decoder approach would have applications in other areas too.

(Zakir Hossain et al., 2018) have discussed the pros and cons of various data sets and techniques used by researchers. CBIR stands for Content-based Image Retrieval is an image indexing method with varied applications that include biomedicine, commerce, the military, education, digital libraries, and web searching.

Figure 2.1 shows the various methods and techniques of image captioning systems.

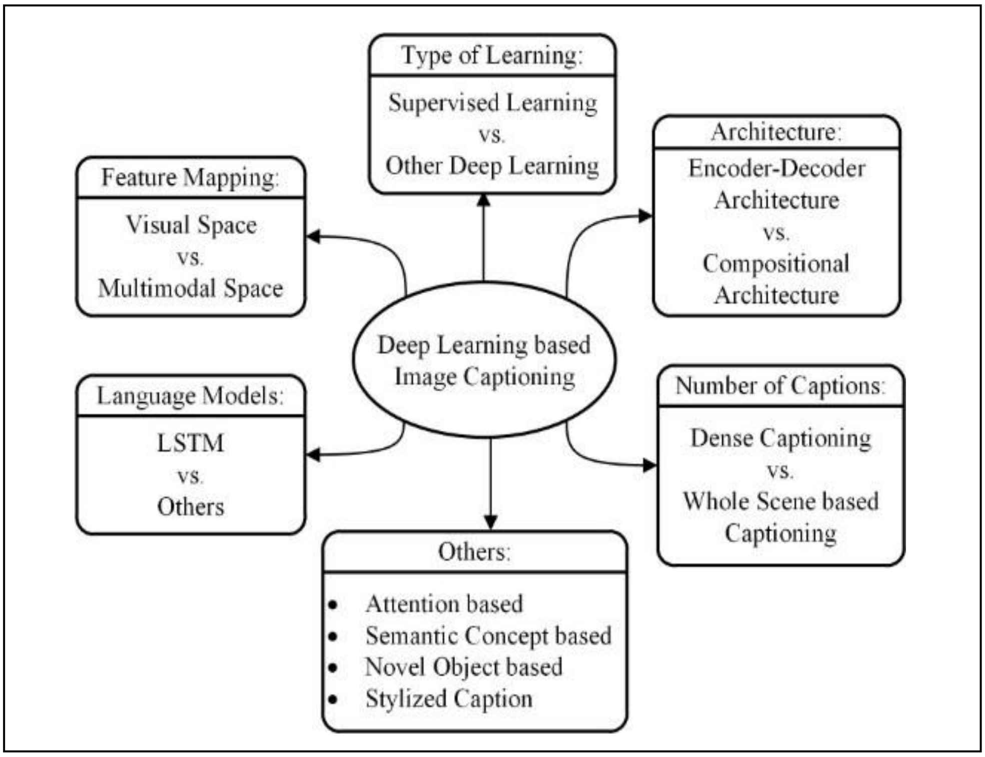


Figure 2.1 **The taxonomy of deep learning-based image captioning system** (Zakir Hossain et al., 2018)

2.3 Discussion

The quest to accurately index images is the main aim of CBIR, there are many uses of such descriptive imagery in the field of commercial applications. Social media platforms like Flicker, Twitter, Facebook, Instagram make use of image captioning systems through the use of Artificial Intelligence. The quality of image detection is based on the system's ability to grasp the image features through the use of traditional machine learning models and deep learning methods. Some researchers have used SVM for image detection where feature extraction takes place from input data which is passed on to a classifier which in turn classifies an object. However, this method has many challenges as extracting all features from diversified and large datasets results in errors. The complexity of videos and real-world data tends to give erroneous captions. (Bernardi et al., 2016; Kumar and Goel, 2017) Deep learning methods are more preferred by researchers and various methods of deep learning are used by researchers.

- Template-based Image captioning
- Retrieval-based image captioning
- Novel image caption generation

Deep learning methods can further be classified as below:-

1. Visual space-based
2. Multimodal space-based
3. Supervised learning
4. Other deep learning

5. Dense captioning
6. Whole scene- based
7. Encoder-Decoder Architecture-based
8. Compositional Architecture-based
9. Long Short-Term Memory
10. Language model-based
11. Others language model-based
12. Attention-Based
13. Semantic concept-based
14. Stylized captions
15. Novel object-based image captioning

Image captions are frequently generated using supervised learning, reinforcement learning, and GAN-based approaches. In supervised learning-based approaches, both visual and multimodal space can be exploited. The most significant distinction between visual and multi-modal space is found in mapping. Methods based on visual space perform explicit mapping from visuals to descriptions. Multimodal space-based techniques, on the other hand, include implicit vision and language models. Encoder-Decoder architecture-based, Compositional architecture-based, Attention-based, Semantic concept-based, Stylized captions, Dense image captioning, and Novel object-based are the several types of supervised learning-based methodologies.

For creating image captions, Encoder-Decoder architecture-based methods use a simple CNN and a text generator. Approaches that focus on distinct prominent areas of the image, such as attention-based image captioning, outperform encoder-decoder architecture-based methods.

Image captioning systems based on semantic concepts selectively focus on different sections of the image and can produce semantically rich descriptions. Image captions based on regions can be generated using dense image captioning algorithms. Image captions that are stylized communicate a variety of feelings such as romance, pride, and shame.

Image Captioning systems based on GAN and RL may create a wide range of captions. MSCOCO, Flickr30k, and Flickr8k are some of the most often used datasets for image captioning.

The MSCOCO collection is quite vast, and all of the photos in it contain several captions. The Visual Genome dataset is mostly used for picture captioning by region.

Image caption performance is measured using a variety of evaluation parameters. BLEU, ROUGE, METEOR and SPICE are the common metrics used for performance evaluation.

2.4 *Summary*

The review of papers suggests the use of modern deep learning techniques along with NLP techniques for the use in image captioning. This area has a wide range of future applications from usage in social media, military intelligence to advanced robotics. The various deep learning methods used for image captioning is studied to select the relevant and most applicable method for this study. The different types of datasets available for research are also explored based on the accessibility and usage by researchers. The evaluation metrics for performance evaluation is also studied for using the methods and evaluating them in the final results.

3.1 *Methods*

The current work is to compare models based on their accuracy for image captioning. State of the art models are selected and explained with the rationale for their selection. Models that are already trained on huge datasets are preferred which tend to give better scores. Pre trained transfer learning models also reduce computational time and costs thereby increasing their efficiency. Each comparison of models brings about their strengths, ease of access and performance. Evaluation metrics that bring uniform assessment on accuracy parameters are also enumerated and would be used in subsequent chapters along with the results.

3.2 *Research Methods*

The study has chosen to explore the below methods and evaluate their performance for image classification and each methods cover their reason for inception, improvements in subsequent versions and popular usage amongst researchers. The below is a model that shows how an image captioning model works.

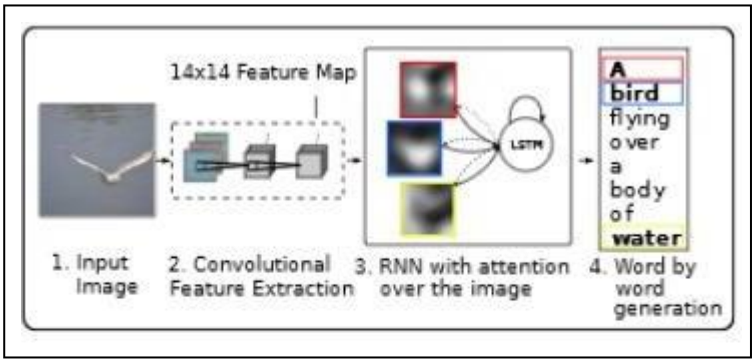

14x14 Feature Map
LSTM
A
bird
flying
over
a
body
of
water
1. Input Image
2. Convolutional Feature Extraction
3. RNN with attention over the image
4. Word by word generation

Figure 3.1: CNNs and RNNs interactions leading to image captioning (Xu et al., 2015)

3.2.1 *Inception V3*

This method was proposed way back in 2015 in a paper titled “Rethinking the Inception Architecture for Computer Vision” by Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, and Jonathon Shlens. The method created better computational efficiency than its inception network predecessors. Inception Networks (GoogLeNet/Inception v1) have been shown to be more computationally efficient than VGGNet, both in terms of the amount of parameters created by the network and the cost incurred (memory and other resources). If an Inception Network is changed, special care must be taken to ensure that the computational advantages are not lost. As a result, due to the uncertainty of the new network's efficiency, adapting an Inception network for different use cases becomes a problem. Several strategies for improving the network have been proposed in an Inception v3 model to loosen the constraints for easier model adaption. Its architecture has the below features: (Szegedy et al., n.d.; Suresh and Keshava, 2019; A Guide to ResNet, Inception v3, and SqueezeNet | Paperspace Blog, 2021; Google AI Blog: Improving Inception and Image Classification in TensorFlow, 2021)

- Factorised Convolutions – checks network efficiancy by reducing the number of parameters involved in a network

- Smaller Convolutions – leads to faster training due to the reduced number of convolutions
- Asymmetric convolutions – the number of parameters would be slightly higher than the symetric convolutions
- Auxiliary classifier – used as a regulizer
- Grid size reduction – reduces cost as operations are pooled together

The neural network has two parts one is a feature extraction part and the other is the classification part. The CNN does the feature extraction as a part of its architecture. The next part of the CNN of Inception V3 is SoftMax layers which are fully-connected to the neural network. The below image from Google AI shows the architecture in a schematic diagram.

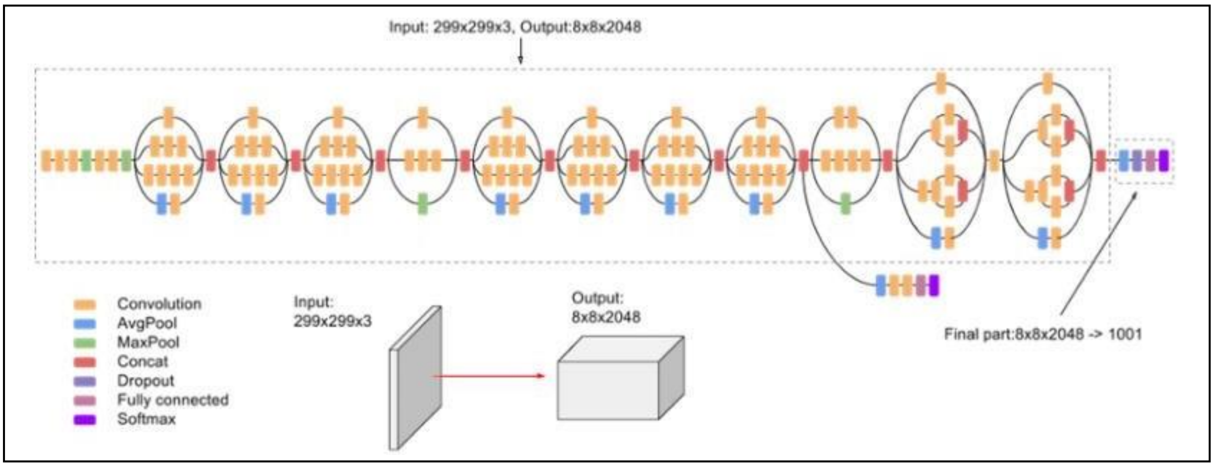


Figure 3.2 Inception V3 - Schematic diagram (Google AI Blog: Improving Inception and Image Classification in TensorFlow, 2021)

3.2.2 *Resnet 152*

(He et al., 2016; Suresh and Keshava, 2019; A Guide to ResNet, Inception v3, and SqueezeNet

| Paperspace Blog, 2021) A team from Microsoft devised this method to overcome the challenges of using deep learning techniques for image classification which were prone to overfitting and where often time consuming. All neural network architectures are not equally easy to optimize. There was need to also address the degradation problem, which occurs when increase in depth leads accuracy to saturate and further degrades rapidly. These challenges were overcome with the usage of residual mapping techniques. Residual networks fit the stacked layers in residual mapping instead of hoping some stacked layers would directly fit the desired underlying mapping. Skip connections, comparable to gated recurrent units, and extensive batch normalization were developed. Resnet won the ICLR 2015 competition exceeding human capabilities.

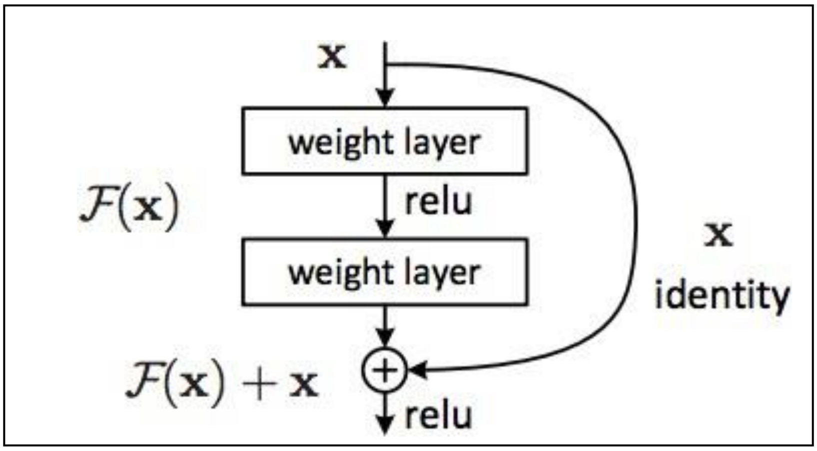

x
weight layer
relu
weight layer
$\mathcal{F}(\mathbf{x})$
x
identity
$\mathcal{F}(\mathbf{x}) + \mathbf{x}$
relu

Figure 3.3 **Residual network architecture (He et al., n.d.)**

3.2.3 *VGG 16*

It's a type of convoluted network only, the numerical 16 indicates the number of layers the architecture comprises of, in this VGG network its 16. This was created with the idea of stacking up layers to form very deep convolutional networks. Its trained-on millions of images from the 'Image net database'. This technique won the best award for the "The ImageNet Large Scale Visual Recognition Challenge 2014". The fully connected nodes size up to 533 MB, and can classify images into 1000 object categories, such as animals, objects like pens and pencils. The input size is 224-by-224.

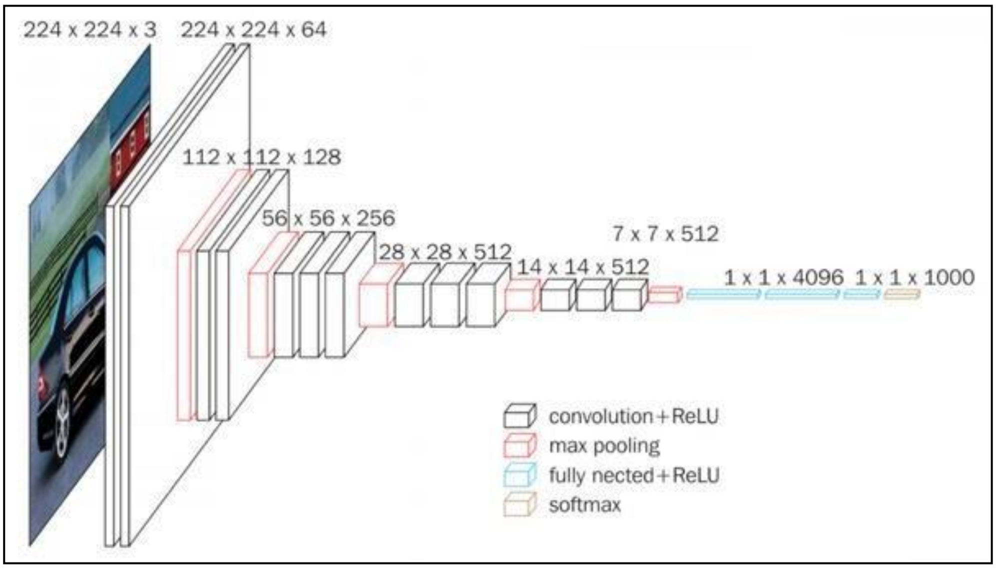


Figure 3.4 **Architecture of VGG 16 (Banerjee et al., n.d.)**

3.2.4 *VGG 19*

Inspired with the success of VGG16 researchers increased the number of layers to 19 it increased accuracy and was a deeper convolutional network than the previous VGG 16. VGG19's input is a 224 224 RGB picture with a fixed size. It has 19 layers, including 16 convolutional layers and three fully connected layers, as well as max-pooling to minimize volume size and a SoftMax classifier after the last fully connected layer. It performs slightly better due to bigger memory than its predecessor. (He et al., 2016, n.d.; Suresh and Keshava, 2019)

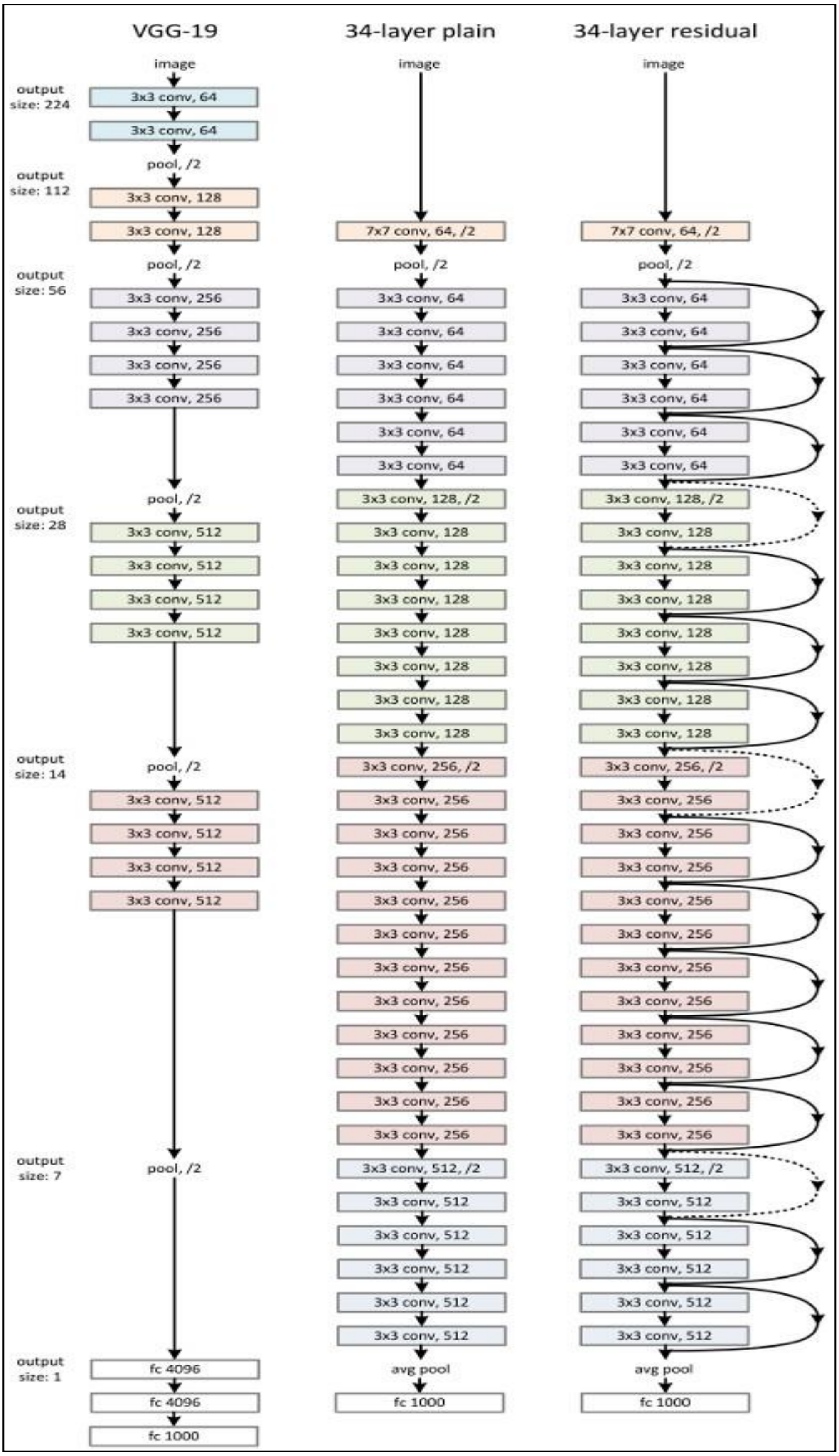


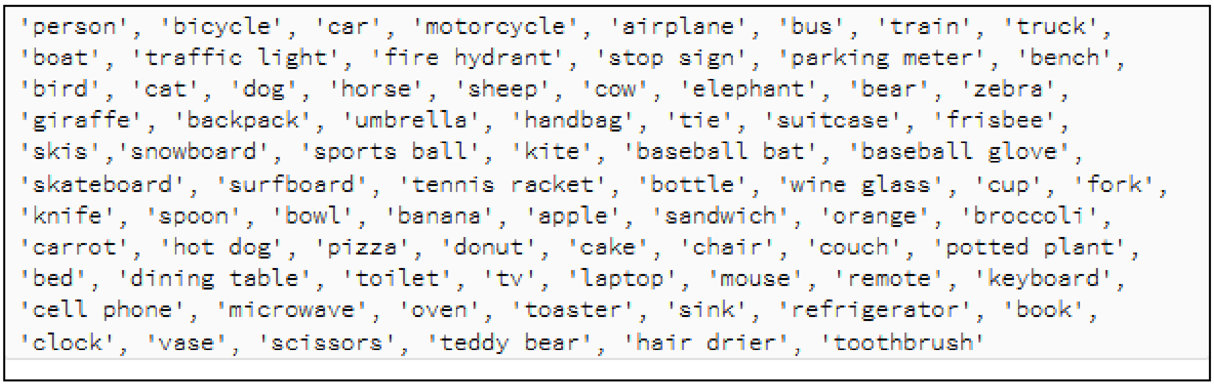
Figure 3.5 **Architecture of VGG 19**

3.3 *Data Set Description and Process Flow*

(Lin et al., 2014; Karpathy and Fei-Fei, 2017; Du et al., 2018; Yan et al., 2018) Tech majors Microsoft and Google are significantly involved in image classification and captioning space. The MS Coco data sets and OID (Open Images Data sets) are widely acclaimed and used by machine learning researchers. This study will make use of the MS Coco data set for comparison of image classification models. The data set is published by Microsoft with the goal of advancing image recognition. The data set is licensed under Creative Commons Attribution 4.0 License which lets users to also use it commercially if the original creator Microsoft is duly acknowledged. It has become tremendously popular for captioning, segmentation and large- scale object detection due to distribution rights, remixing and tweaking rights. COCO is abbreviated as Common Objects in Context. Its often used as a benchmark dataset to determine comparative performance of machine learning algorithms. The data set contains high quality visual images for computer vision.

The main reasons for choosing COCO data set for this study are the below:

1. 1.5 Mio object instances
2. Super pixel stuff segmentation
3. Recognition in context
4. Object segmentation with detailed instance annotations
5. Over 60% of the images are labelled which provides labelling for over 200000 images
6. Coco classes are object categories that number 80 detailing individual instances
7. Coco Stuff are objects with no clear boundaries panoptic: full scene segmentation, 91 stuff categories exist in the data set which provide significant contextual information
8. 5 captions per image
9. Large data set containing 328000 images
10. DensePose annotations have been applied to over 39,000 photos and 56,000 person instances, with each labelled person receiving an instance id and a mapping between image pixels attributable to that person's body and a template 3D model. The annotations are publicly available only for training and validation images.



```
'person', 'bicycle', 'car', 'motorcycle', 'airplane', 'bus', 'train', 'truck',
'boat', 'traffic light', 'fire hydrant', 'stop sign', 'parking meter', 'bench',
'bird', 'cat', 'dog', 'horse', 'sheep', 'cow', 'elephant', 'bear', 'zebra',
'giraffe', 'backpack', 'umbrella', 'handbag', 'tie', 'suitcase', 'frisbee',
'skis','snowboard', 'sports ball', 'kite', 'baseball bat', 'baseball glove',
'skateboard', 'surfboard', 'tennis racket', 'bottle', 'wine glass', 'cup', 'fork',
'knife', 'spoon', 'bowl', 'banana', 'apple', 'sandwich', 'orange', 'broccoli',
'carrot', 'hot dog', 'pizza', 'donut', 'cake', 'chair', 'couch', 'potted plant',
'bed', 'dining table', 'toilet', 'tv', 'laptop', 'mouse', 'remote', 'keyboard',
'cell phone', 'microwave', 'oven', 'toaster', 'sink', 'refrigerator', 'book',
'clock', 'vase', 'scissors', 'teddy bear', 'hair drier', 'toothbrush'
```

Figure 3.6 Snapshot of 80 objects present in the MS Coco data set (Lin et al., 2014; COCO - Common Objects in Context, 2021)

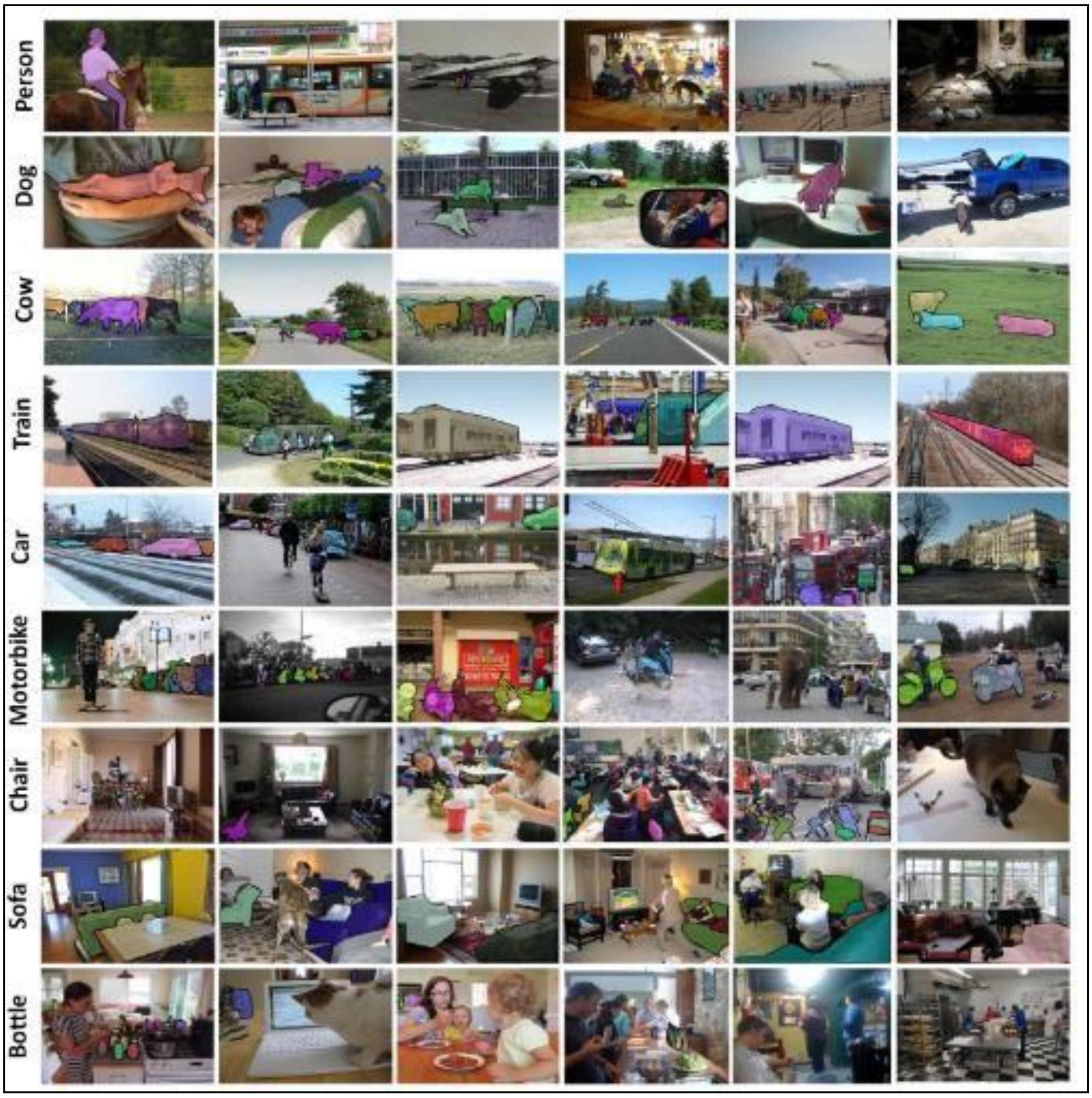


Figure 3.7 Annotated images in MS COCO data set (Lin et al., 2014; COCO - Common Objects in Context, 2021)

3.3.1 *Process Flow*

Below is the proposed process flow and subject to change based on initial runs and logistical challenges faced (if any)

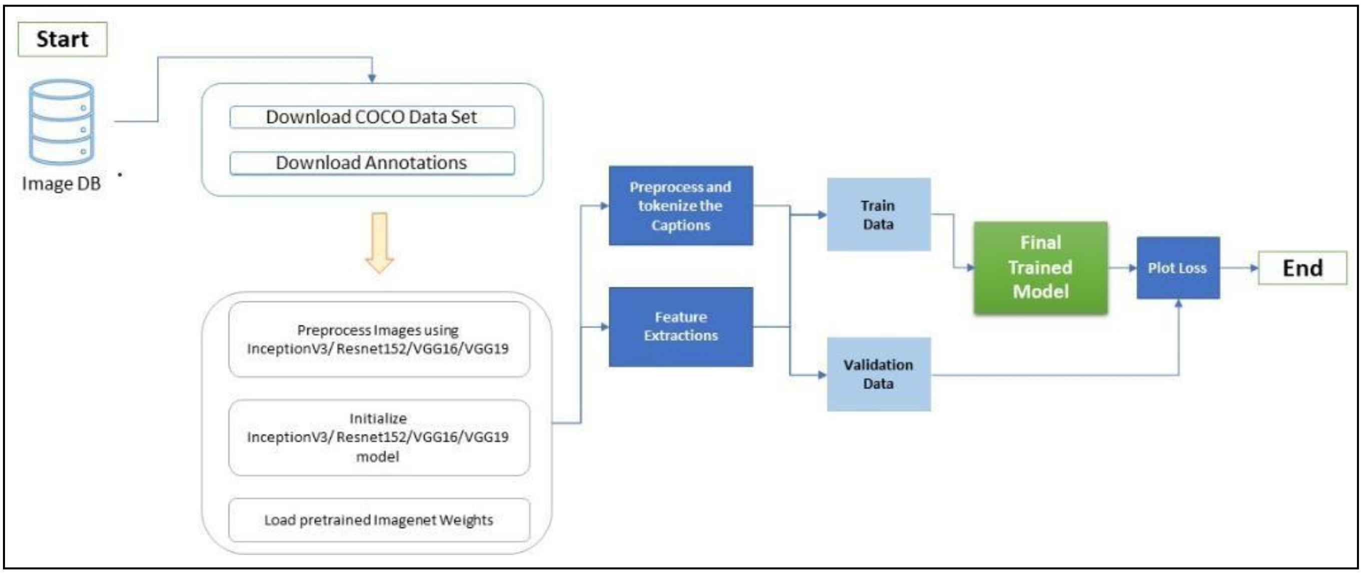

Start
Image DB
Download COCO Data Set
Download Annotations
Preprocess Images using InceptionV3/ Resnet152/VGG16/VGG19
Initialize InceptionV3/ Resnet152/VGG16/VGG19 model
Load pretrained Imagenet Weights
Preprocess and tokenize the Captions
Feature Extractions
Train Data
Validation Data
Final Trained Model
Plot Loss
End

Figure 3.8 Process Flow Diagram

3.4 *Evaluation Metrics*

The suggested approach has the potential to accurately represent visual content in an image while also improving precision with all outcome assessment indicators. Tourist attractions and celebrities are out of the question. This method's generated image descriptions are evaluated based on metrics. The results of the evaluation measurements on image captioning techniques have been excellent. Though, in terms of image semantics assessment, there is still a significant difference between the matrix assessment on the method and the human decision to assess.

The report will concentrate on weighing the advantages and disadvantages of the generation result based on the degree of similarity between the caption and reference sentences.

The five-calculation metrics BLEU, METEOR, ROUGE, CIDEr, and SPICE are the most widely used. BLEU and METEOR are machine translation-based indicators, ROUGE is based on text abstraction, and CIDEr and SPICE are image captioning-based indicators.

3.4.1 *BLEU*

The BLEU algorithm, which is built on the n-gram precision, is widely used to evaluate image annotation results. The BLEU metric works on the principle of calculating the interval between the measured and reference sentences. Where the caption is closest to the duration of the comparative sentence, the BLEU approach attempts to give a higher ranking.

Advantages of BLEU: -

1. Its language independent
2. Easy comprehension
3. Computational ease
4. Scores lie between [0,1] indicative of quality captions with higher scores

3.4.2 *ROUGE*

ROUGE is a machine-learning-based validation standard for text summarization algorithms. ROUGE-N, ROUGE-L, and ROUGE-S are the three-evaluation metrics. ROUGE-N calculates a basic n-tuple recall for all comparative statements relying on the given sentence to be evaluated: ROUGE-L measures recall using the largest general series (LCS). ROUGE-S computes recall using skip-bigram co-occurrence statistics between reference and projection text descriptions.

3.4.3 *CIDEr*

CIDEr is a special technique that can be applied for image captioning. It determines the term frequency inverse text frequency (tf-idf) for each n-gram to indicate the degree of agreement in image captioning. According to studies, the CIDEr and human consensus match is superior than other evaluation criteria.

3.4.4 *METEOR*

METEOR is dependent on a harmonic mean of unigram precision and recall, with recall getting a greater weight than accuracy. It differs from the BLEU in that it is found not only in the whole sequence, but also at the sentence and segmentation stages, and it has a positive association with human judgement.

3.4.5 *SPICE*

SPICE evaluates the quality of image captions by transforming the generated description and reference sentences into “scene graphs,” which are graph-based semantic representations.

The scene graphs extract lexical and syntactic information by natural language and represent the image's objects, characteristics, and relationships.

4.1 *Implementation*

In this chapter we look at how to implement and run algorithms on the chosen data set. The previous chapters discussed the review of researchers solving image captioning through the various use of algorithms, the methodology, data sets and the evaluation metrics. The implementation is decided with the cross section of the MS Coco data set. The data set contains more than 300000 images and is extremely a large data set to handle with limited resources.

The cross section of the data set consists of 50000 images that are randomly chosen This makes it easier to handle with the given resources at the researcher namely time, costs and processing capacity.

4.2 *Resources*

The computing systems used for this study is a Dell Latitude 3410 business laptop, is a 10th generation system with a core i7 processor, with computing speeds of 1.8 GHz has 8 giga bytes of RAM (DDR4, SDRAM) having 256 GB of internal storage. The initial runs couldn’t handle the data size of the files with multiple process kills. This led to the making use of external resources which are nimble box a full stack MLOps platform for ML researchers. Subscription to Google Colab Pro was also taken its CPU is Intel® Xeon 2.3 GHz, and an external GPU which is subscription mode only and runs on Tesla P100-PCIE-16GB which could handle all the runs with lesser number of kills and downtime.

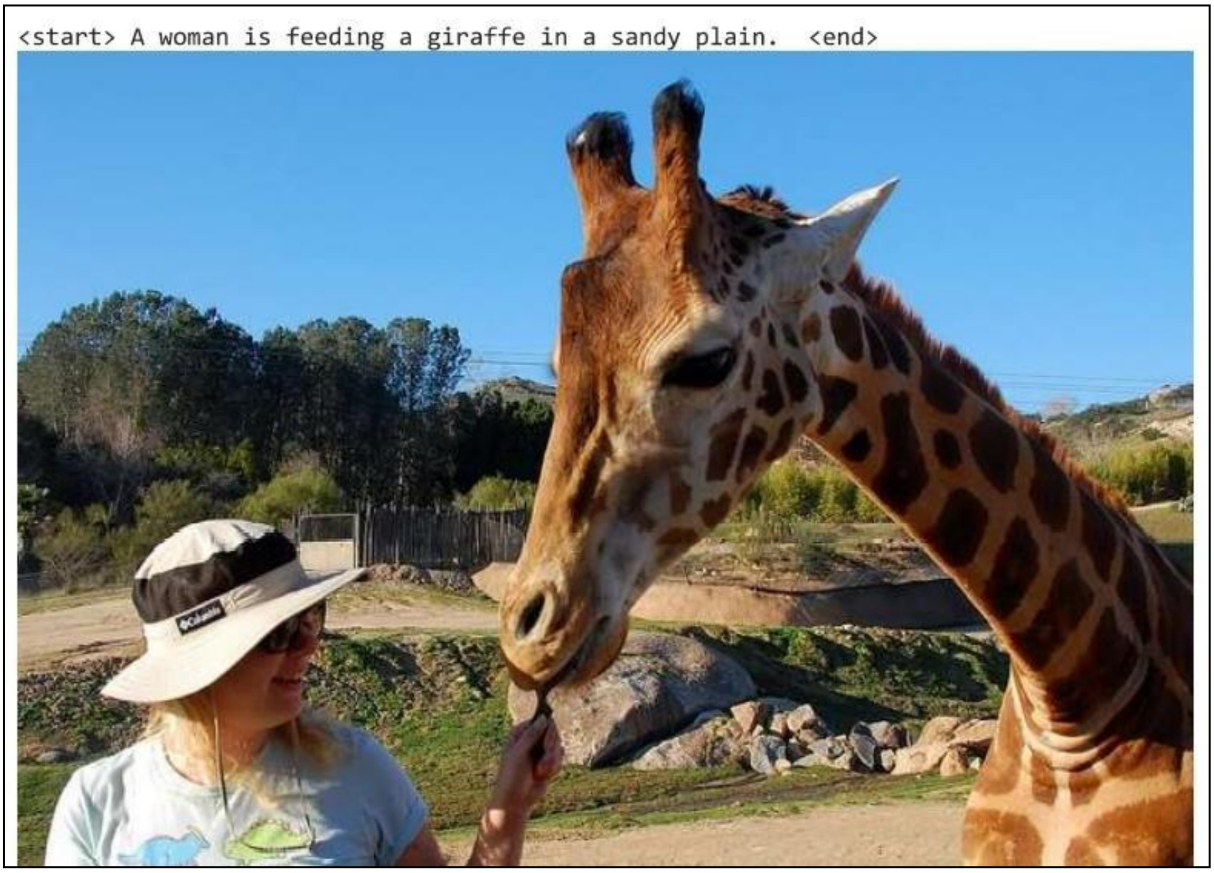
4.3 *Packages*

Standard python packages have been used for this study and run on the latest Python version

3.8.0. (version) The study requires generating plots using matplotlib pyplot libraries for calculating loss plots and other compositions of the data sets and preliminary exploratory data analysis. Keras (version) is used as its open source and needed for the usage of neural networks to analyse images of the data set. The API is made use of as a high-level wrapper and run on TensorFlow which is a great enabler to run machine learning models on desktops, PCs and cloud. NumPy library for mathematical and linear algebraic functions. NLTK packages for natural language programming tasks.

4.4 *Data Analysis and Pre-processing*

Cross section of the data set was used for this study which totalled 50k images, loading the data ranged from 530 to 1170 seconds. All captions for the same image ID were grouped together. Sample image of the loaded data set with caption is shown in fig 4.1. Removal of stop words, removal of punctuation marks was carried for select methods like resnet, vgg16 and vgg19. Tokenisation which is to separate words, characters or sub words which makes it easier to process and the machines to comprehend was also carried out for resnet, vgg16 and vgg19 models. The convoluted neural networks which are advanced, and state of the art are also able to handle lot of pre-processing tasks themselves without the need for manual intervention and inception v3 was also able to handle the pre-processing tasks by itself preparing the data for further processing to generate captions for images.



```
<start> A woman is feeding a giraffe in a sandy plain.  <end>
```

Figure 4.1 Image loaded with caption in the data set

4.5 *Model Implementation*

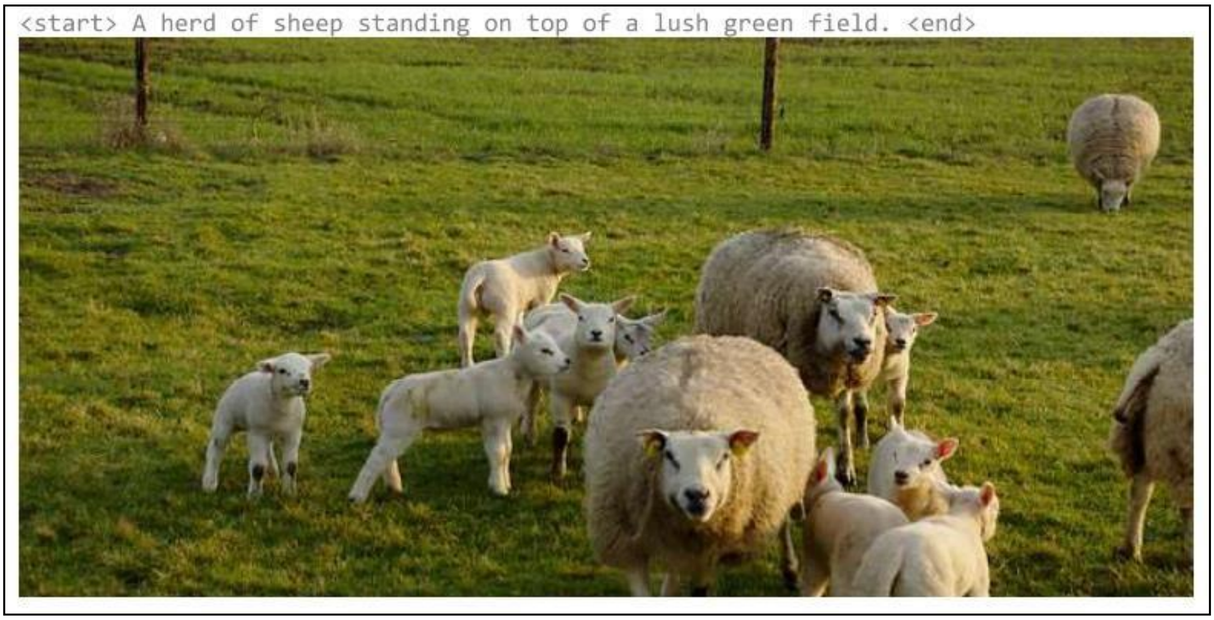
Models' inception V3, Resnet, VGG 16 and VGG 19 are implemented for this study after the initial tasks of loading the datasets grouping annotations for the same image id together and making random checks on the datasets. The subsequent sections will talk about the model implementation of each model, the time taken and the output shared. The results comparison of each model will be discussed in the next chapter titled as "Results". The model runs faced considerable limitations due to the limited resources and time. Many runs exceeded 12 hours of run time on demand, allowed under the subscription package of Google Colab Pro which led to multiple kills and took more than 50 hours of run time for running models and executing them successfully.

4.5.1 *Inception V3*

The limited resources and time let to the trial of running Inception V3 in two different attempts with different data sizes. The initial run was completed using 9K images and after the successful run, the same model was run with a total data set size of 50K images which took the maximum time of 50 hours of multiple run times to generate the scores. The below sub sections will deal with individual implementation of both the runs of the different data sets of images. The run time and implementation time had a plausible variance due to the change in the number of images.

4.5.1.1 *Inception V3 9k*

The first trial of implementing the model inception V3 was undertaken with 9k images. Packages were first imported into python which are matplotlib for plotting, NumPy, keras, tensorflow, collections, random, os, time, json, jury which would be needed for further implementation of the models. The data consisted of importing the images file and the annotations file. The annotations file took 17 seconds to load and the images file took 1247 seconds files to fully load the files on the Google Colab Pro external severs. 9K images with 5 image captions totals to 45K examples for this analysis. A random image from the set of 9K images is shown in **Figure 4.2**.



```
<start> A herd of sheep standing on top of a lush green field. <end>
```

Figure 4.2 **Random image from the 9K data set along with its caption**

InceptionV3 will extract features from the very last convolutional layer after becoming pre- trained on ImageNet to classify each image. Inception V3 requires pre-processing of images in the required format to process them. Below are the two pre-processing steps needed for inception V3

1. Normalization of images so that images are in the pixel ranges of -1 to +1, which is needed for training images
2. Each image is resised to 299px by 299px

Images were present with no captions and captions were also present in the data set with no image ids. Such captions and images were removed and only images with captions were sorted and chosen for further analysis. Annotations were tokenised limiting the vocabulary to the top 8K words. Adam optimiser algorithm was used to increase the learning rate and minimise loses. The inherent reasons for its implementation and usage is due to its ease of implementation, computational efficiency, low memory requirements, invariance to diagonal rescale of the gradients, better for dealing with noise gradients like image data sets and low tuning of hyperparameters. The feature selection and recognition were as below and represents the vector shape.

- 2048 features (non attention)
- 64 attention feautures

The loss function flattened at 0.2 at 250 epochs as shown in figure 4.3. The training split was done at 70:30 split and also on 80:20 split. The output consisted of txt files with predicted and real captions as shown in figure 4.4.

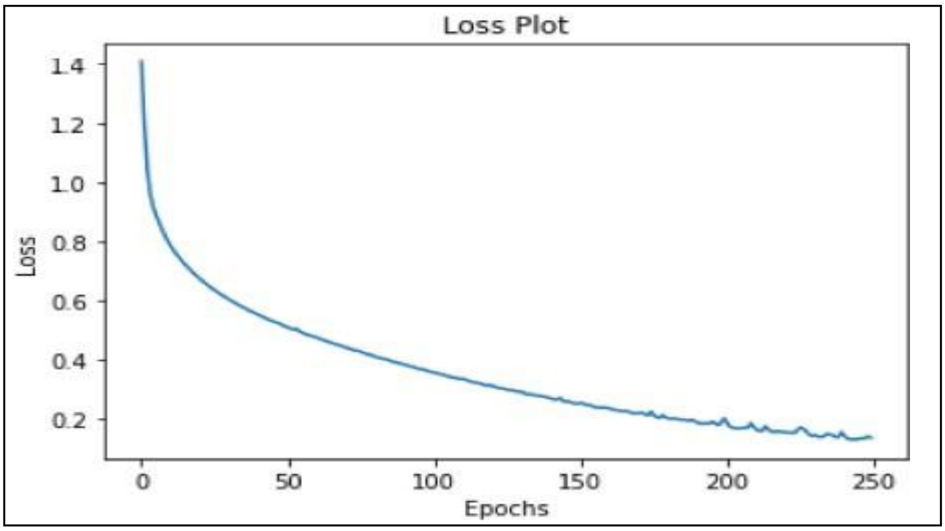
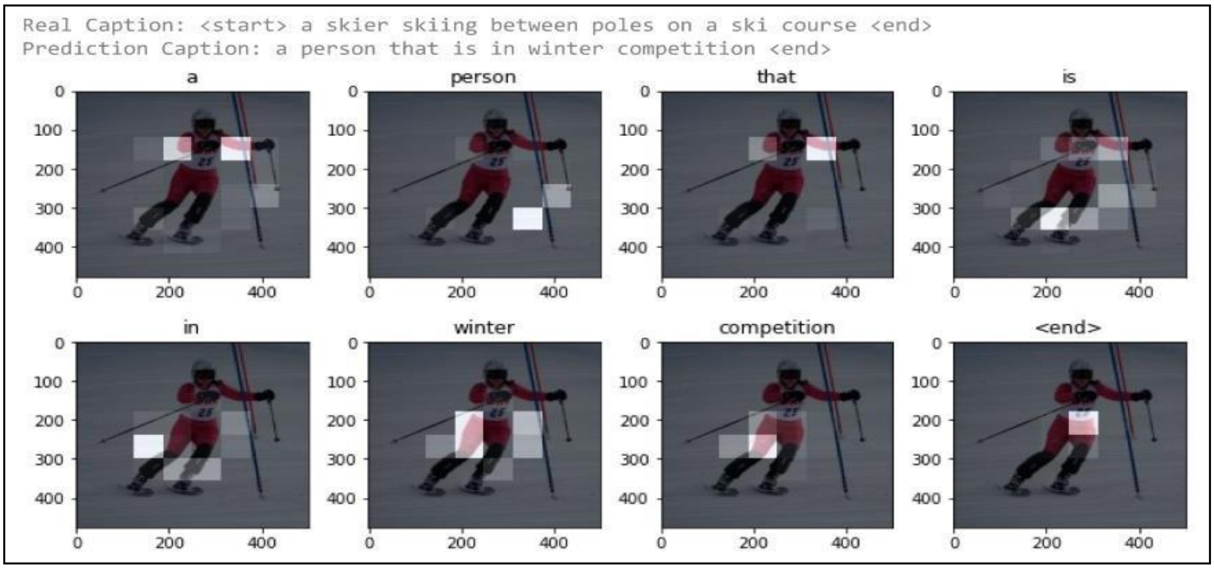


Figure 4.3 **Output Loss plot for Inception V3 9K**



```
Real Caption: <start> a skier skiing between poles on a ski course <end>
Prediction Caption: a person that is in winter competition <end>
```

Figure 4.4 **Output with predicted and real caption for a chosen image**

4.5.1.2 *Inception V3 50k*

The output yielded with 9K images successfully led to further trials with a larger data set with the size of 50K images. All the previous packages installed were again imported for another run matplotlib for plotting, numpy, keras, collections, random, os, time, json, jury and tensorflow. The annotations file took 8 seconds to load and the images file took 541 seconds to fully load in the external Google Colab Pro server. All images with the same image ID were grouped together. Images had 5 captions and with 50K images totalled to 250K examples for the cross section of the data set.

Pre-processing steps like normalizing images pixel range between -1 to 1 were carried out to match the training requirement for Inception V3 and resizing of images to 299px by 299px as done on the previous implementation for the 9K run. Images were also sorted to select images which had captions and reject images with no captions or captions with no images to further process the analysis. Captions were tokenised with the top 8K selected for vocabulary for further training. Vector shape was again (64, 2048) with the first value 64 representing attention features shape and the later features (without attention). Inception V3 works on encoder, decoder mechanisms where the vector is passed to CNN Encoder which is singular fully connected layer and the decoder RNN (GRU) predicts the next word for the image. The test- train split was carried out on both 70:30 and 80:20 splits. This was run on a total of 100 epochs with each epoch taking 1006.25 seconds to run totalling to 100625 seconds for all the epochs. However, the total time taken for all the 100 epochs to run were 50 hours due to on demand GPU limitations with code kills in excess of 12 hours. Adam X optimiser was used to increase efficiency of the learning rate and minimise loses. The loss plot flattened at 0.327107 as shown in figure 4.5. The final output was generated in TXT format for predicted and actual captions as per the figure 4.6. Scores were also generated which will be discussed in the next chapter.

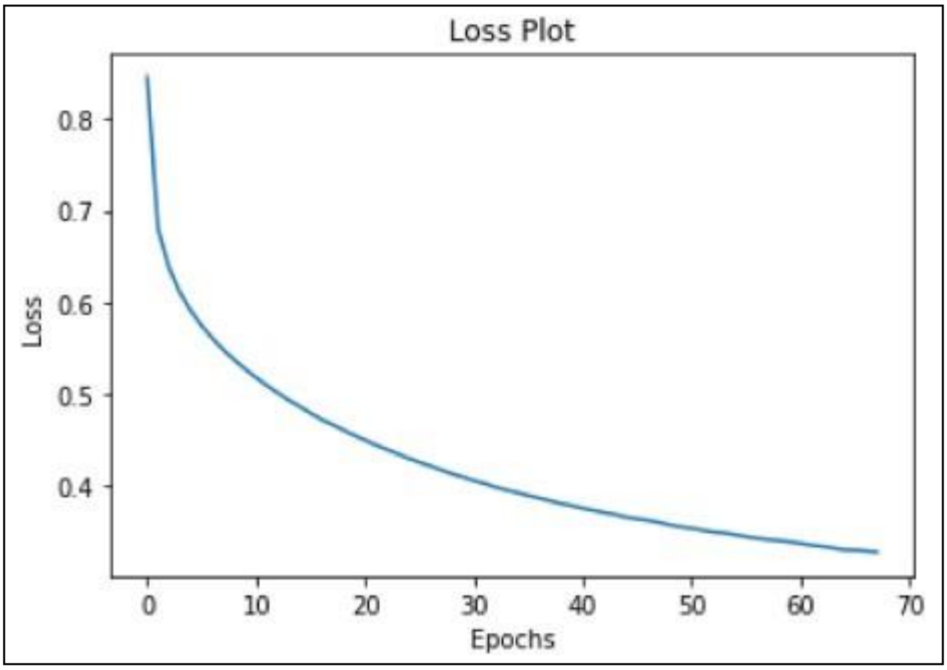
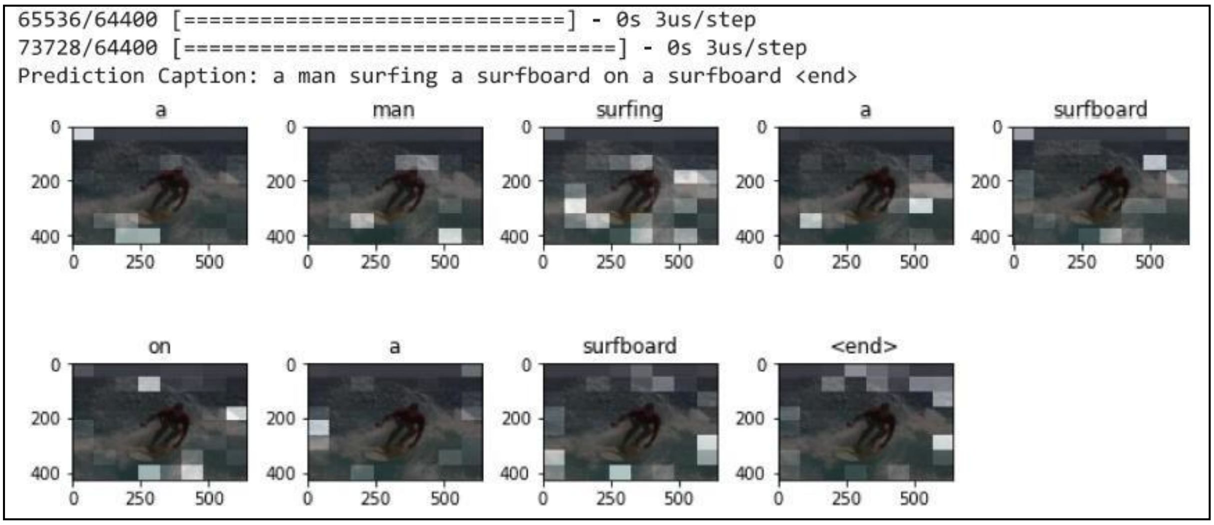


Figure 4.5 **Output Loss plot for Inception V3 50K**

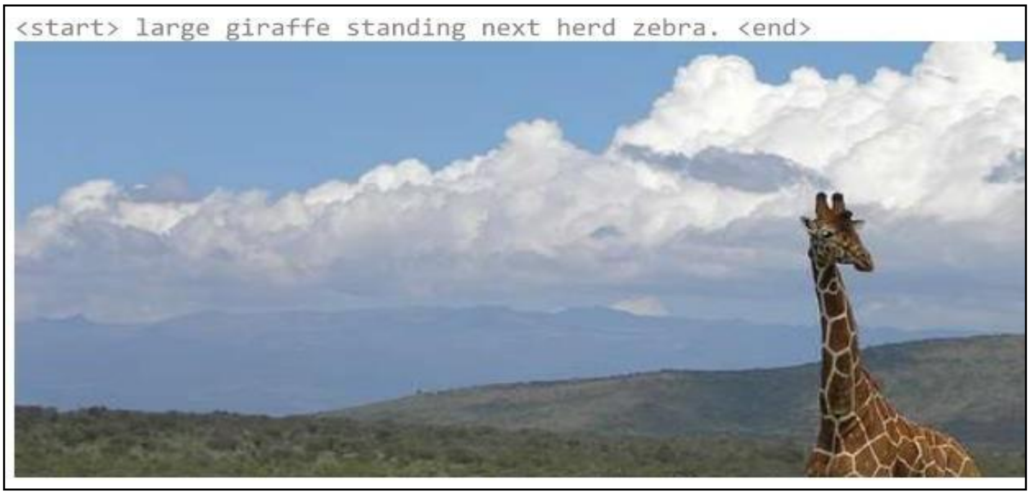


```
65536/64400 [==============================] - 0s 3us/step
73728/64400 [==================================] - 0s 3us/step
Prediction Caption: a man surfing a surfboard on a surfboard <end>
```

Figure 4.6 **Output with predicted and real caption for a chosen image**

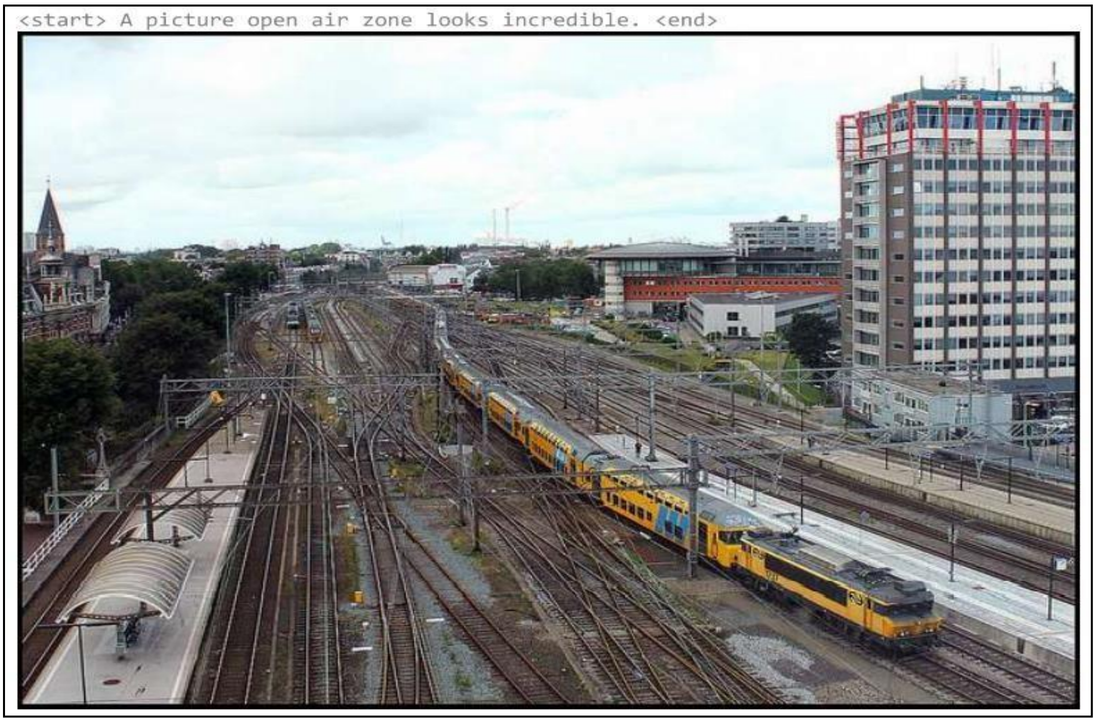
4.5.2 *VGG algorithm version 16 and 19*

This model implementation required importing packages matplotlib.pyplot for plotting, tensor flow, keras, collections, random, os, time, json and jury. The annotations file took 14 seconds to load and the images took 421 seconds due to the lesser image size. A total ok 9K images were selected for implementation with VGG16 and 19 algorithms. Pre-processing steps included, removing stop words, punctuation marks. 9K images with 5 captions had 45K examples for the further analysis. Random image of the cross section of the image date set is shown in figure 4.7 and 4.8. Images were resized 224px to 2244px needed for training requirement requirements running on VGG 16 and 19. Other pre-processing steps included tokenising words and selected top 8K words for total vocabulary. Selection of images with captions and rejecting Image IDs without caption or captions without image IDs.



```
<start> large giraffe standing next herd zebra. <end>
```

Figure 4.7 **Random image from the 9K data set along with its caption for VGG16**



```
<start> A picture open air zone looks incredible. <end>
```

Figure 4.8 **Random image from the 9K data set along with its caption for VGG 19**

The test train split was done at 80:20 split, the feature extraction and the vector shape for VGG 16 is (49, 2048) with 49 representing attention features and 2048 representing features (non- attention). VGG 19 had a vector shape of (196, 2048) with 196 representing attention features and 2048 representing non attention features. The learning rate was 0.0001 using Adam optimiser. The VGG was run on 250 epochs for version 16 and 230 epochs for version 19. The loss plot flattened at 0.140 at 200 epochs for VGG 16 and at 0.1204 with 222 epochs for VGG 19 shown in figure 4.9 and 4.10 respectively. The output with predicted captions was generated into a TXT file and scores were generated which would be discussed in the next chapter.

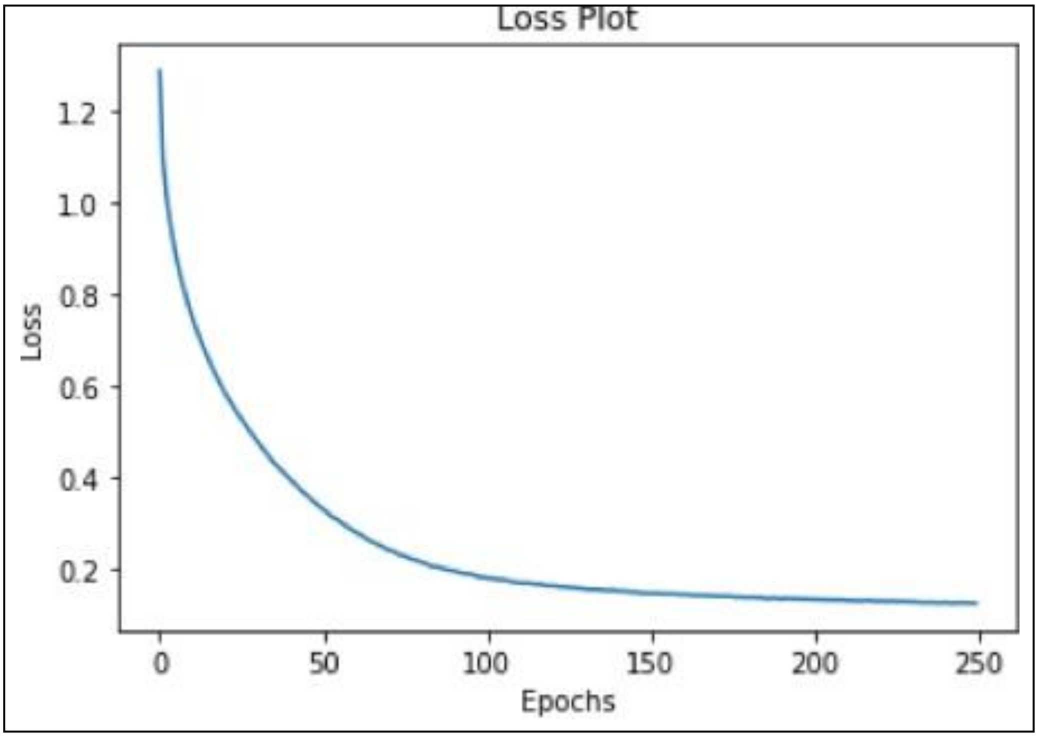


Figure 4.9 **Loss plot for VGG 16**

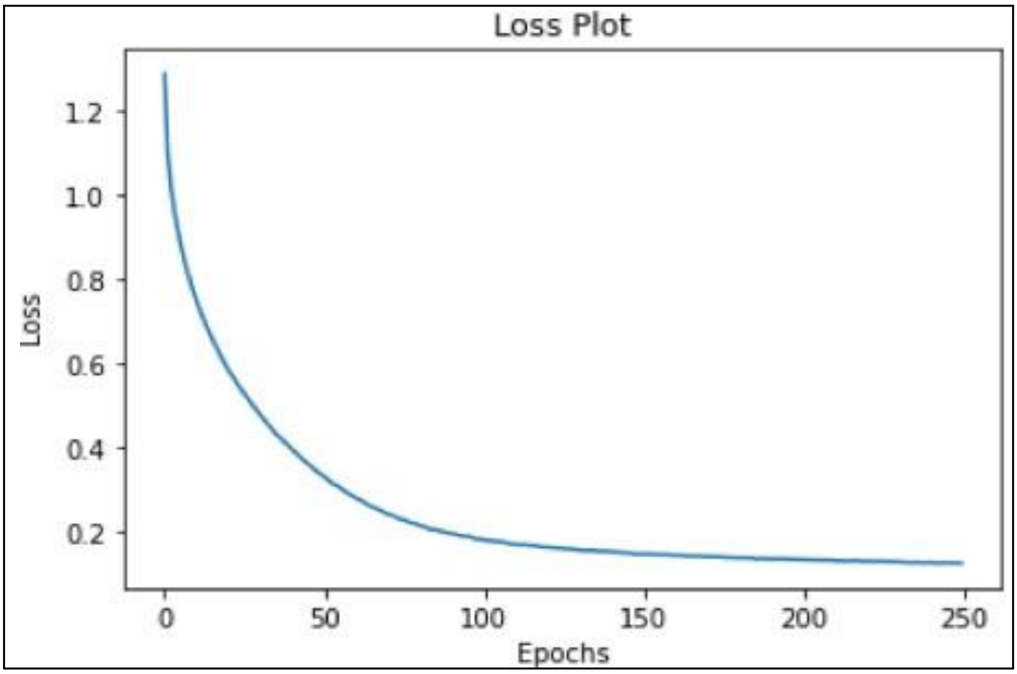
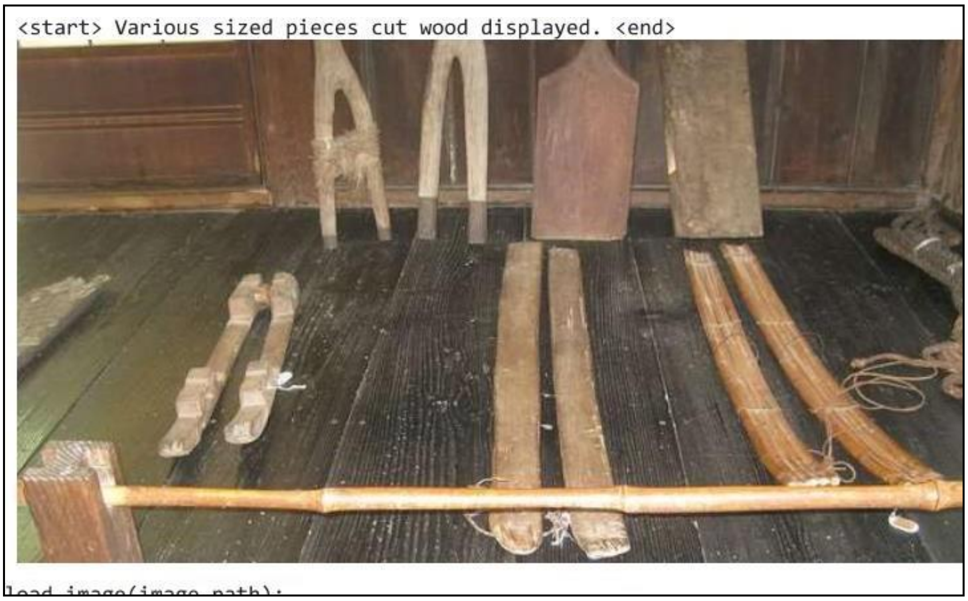


Figure 4.10 **Loss plot for VGG 19**

4.5.3 *Resnet*

The implementation included multiple image captioning algorithms using encoder decoder mechanisms of one which is Resnet. Though its not advanced as InceptionV3 being a residual network, however is much celebrated due to being judged as winner performing better than humans in image captioning competitions held in 2015, this algorithm is chosen for analysis to see and compare the results. Standard packages similar to the ones imported for previous algorithms are imported which are TensorFlow, Keras, collections, random, numpy, os, time, json, jury. The image data set size consists of 9k images. Annotation's file took 8 seconds to load and 539 seconds for the images file to the on-demand Google Colab pro server. Pre- processing steps involved were removing stop words, tokenising, removing punctuation marks. The data set had 9K images with 5 captions each providing 45K examples for further analysis.

Random image from the 9K data set is shown in the figure 4.11, top 8K words were selected for vocabulary and images with captions were only considered for implementation. The maximum length of the caption was 32. The split was done on 70:30 and 80:20 splits. The feature extraction for Resnet 152V2 is (49, 2048) with 49 representing attention features and 2048 features (non-attention) showing the vector shape. Adamax optimiser was used to enhance learning rate and minimise losses, the learning rate was 0.0001. The algorithm was run on 150 epochs and loss plot flattening at 80 epochs around 0.18 as shown in figure 4.12. The output was generated in TXT file along with the accuracy scores.



```
<start> Various sized pieces cut wood displayed. <end>
load_image(image_path);
```

Figure 4.11 Random image from the 9K data set along with its caption for Resnet 152

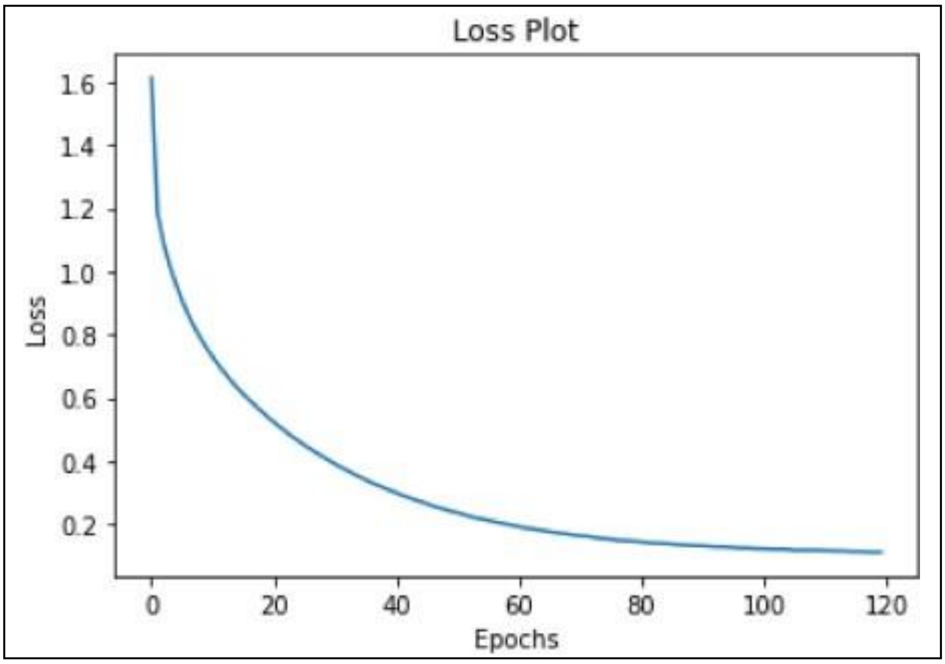


Figure 4.12 Loss plot for Resnet 152

4.6 *Summary*

The model implementation involved choosing the right size of the data sets, selecting the right types of images, pre-processing of images to prepare them for further processing to match the training requirements of the algorithmic models used for image captioning. Each of the chosen models, Inception V3, Resnet 152 V2, VGG 16 and 19 had its own challenges and run times considering the implementation was done on an on-demand subscription package. The Colab Pro GPU kills code beyond runs of 12 hours and some of the models took more than 50 hour and over 100 epochs to complete the runs successfully to generate the predicted captions and the requisite scores for further analysis. The algorithm Inception V3 was run on both 9K images and 50K images making it a much larger data set presenting data handling challenges with regards to resources and time. The other algorithms VGG 16, 19 and Resnet ran successfully and faster than Inception V3 generating scores and predicted captions which will discussed in the next chapter to compare results and accuracy of generated scores. The implementation provides scores both for different data size comparison using the same model and for the same data size using different models.

5.1 *Results*

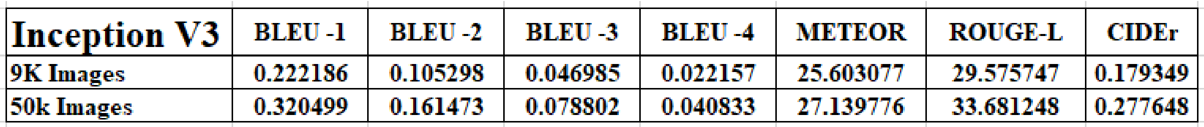
The previous chapters took us through the literature review to better understand the research carried on image captioning by various researchers. The selection of the right data set and implement algorithms that are apt for image caption predictions. This chapter will see the analysis of the select scores for image captioning and evaluate based on the scores generate to understand which algorithms are able to fare better. The predictive power of each algorithm in caption generated will be the basis of this comparative study on image captioning.

5.2 *Model Output*

The model output will be analysed for different data set sizes for Inception V3 and a comparison of the same evaluation metrics for similar dataset sizes. The below table captures the scores for Inception V3 run on two data set sizes for 9K and 50K. Subsequent sub sections presents the results for using Inception V3 for different data set sizes and evaluated its performance based on the increased training data made available.

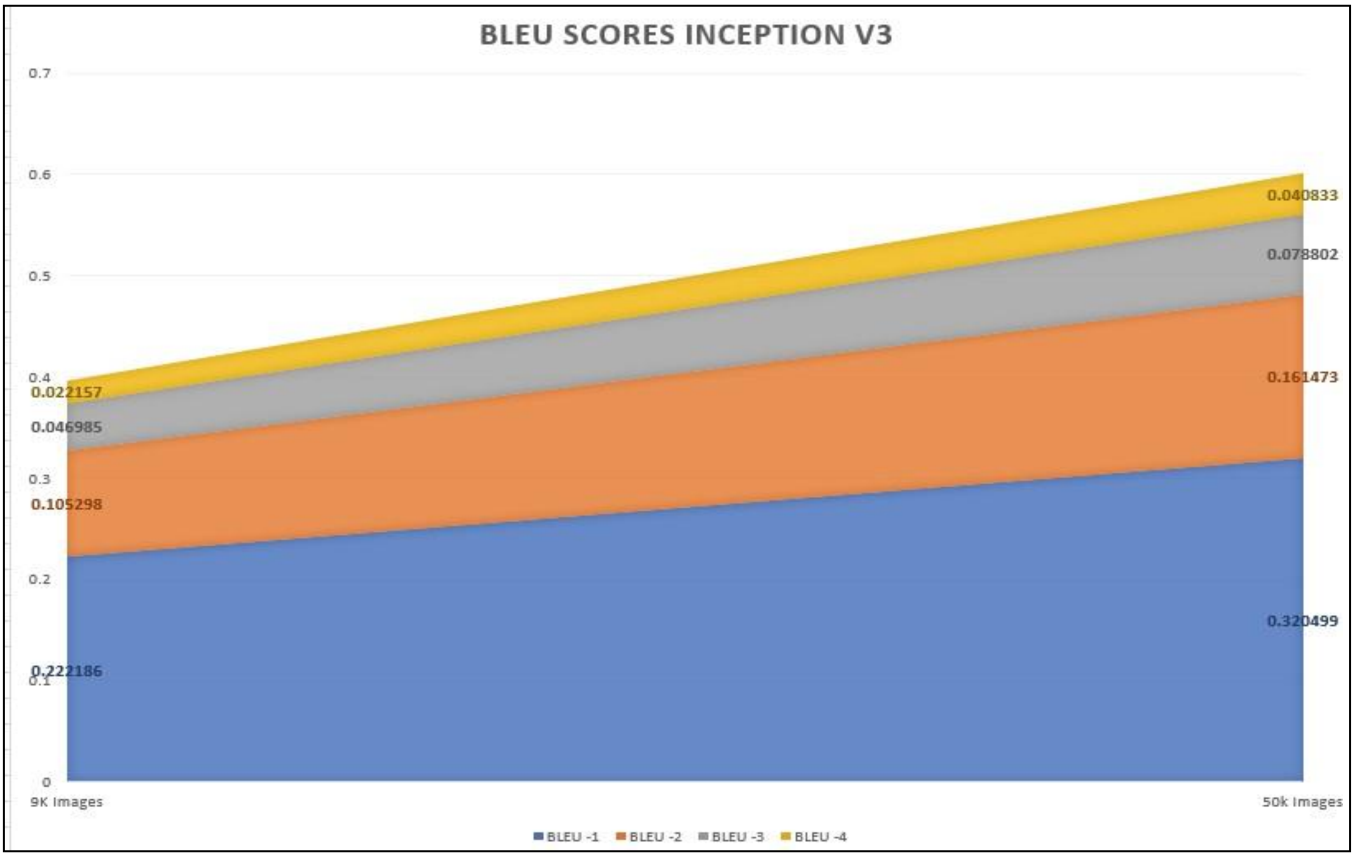
5.2.1 *Comparison of Inception V3 scores on different data set sizes of 9K and 50K*

The scores show considerable improvement across for increased data set sizes, due to the higher number of images for training and considerably boosting their predictive power to generate appropriate captions. The figure 5.1 exclusively shows BLEU scores for all its variants when run for inception V3 algorithms which clearly establishes significant performance for increased data set sizes. Only METEOR and ROUGE show only marginal improvements compared to BLUE scores. ROUGE calculates performance which is much similar to BLEU scores but uses a different n grams variant and calculates recall for predicted and existing captions. METEOR uses n grams with synonym matching and increased data set sizes hasn't largely impacted its performance.



Inception V3	BLEU -1	BLEU -2	BLEU -3	BLEU -4	METEOR	ROUGE-L	CIDEr
9K Images	0.222186	0.105298	0.046985	0.022157	25.603077	29.575747	0.179349
50k Images	0.320499	0.161473	0.078802	0.040833	27.139776	33.681248	0.277648

Table 5.1 **Scores generated using Inception V3 for image data sizes**

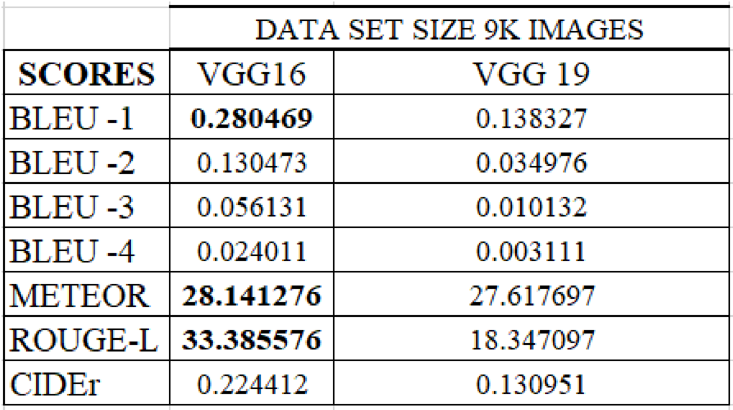


BLEU scores have shown significant performance compared to the smaller data set, BLEU 1 has shown an improvement of

10 percentage points due to the increased data set size. BLEU 4 scores have doubled due to the increased data set size.

5.2.2 *Comparison of scores for VGG 16 and VGG 19 on same data set size*

The VGG algorithm is run on version 16 and 19 to compare it performance across evaluation metrics. VGG 16 has given better performance than VGG 19 on all evaluation metrics, the BLEU -1 score is double the percentage points and can be considered as a better performing. version over the newer version. The newer version has shown poor performance on all BLEU evaluation metrics variants 1-4 with BLEU 2 and 4 shows far worse performance. The METEOR metric is comparable across VGG versions and is the only metric to show such a semblance. Table 5.2 shows the individual scores of both the versions of VGG net. Figure 5.2 highlights the better performance of the older VGG 16 version over its newer variant.



	DATA SET SIZE 9K IMAGES	
SCORES	VGG16	VGG 19
BLEU -1	**0.280469**	0.138327
BLEU -2	0.130473	0.034976
BLEU -3	0.056131	0.010132
BLEU -4	0.024011	0.003111
METEOR	**28.141276**	27.617697
ROUGE-L	**33.385576**	18.347097
CIDEr	0.224412	0.130951

Table 5.2 Comparison of Individual scores generated using VGG 16 and 19

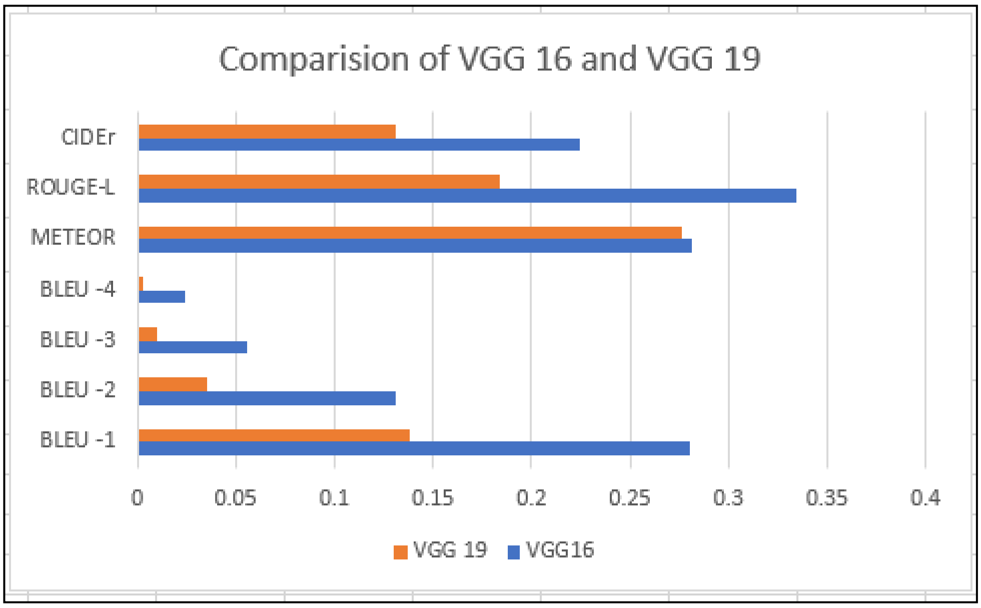
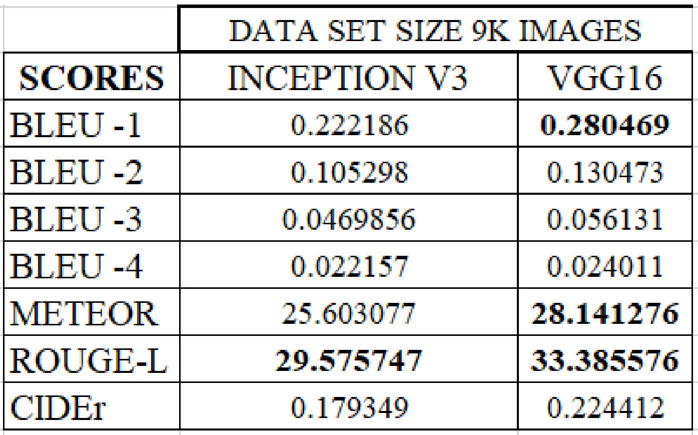


Figure 5.2 Comparison of VGG 16 and VGG 19 run on same data set sizes

5.2.3 *Comparison of scores for VGG 16 and Inception V3*

VGG 16 outperforms all other algorithms on the same data set sizes, for INCEPTION V3. Its performance has fared better than even its newer version VGG 19. Table 5.3 shows the comparison of VGG 16 and Inception V3. BLEU score of 0.28 is best amongst all the BLEU metric variants. METOR and ROUGE show comparable performance, though VGG 16 still outperforms them on all metrics. CIDEr being a tf-idf weighted n gram still lags behind other recall-based metrics like BLEU and more human consensus-based metric. Figure 5.3 captures the overlap of scores for BLUE metric variant 3 and 4 and better performance over inception v3 across all evaluation metrics.



	DATA SET SIZE 9K IMAGES	
SCORES	INCEPTION V3	VGG16
BLEU -1	0.222186	**0.280469**
BLEU -2	0.105298	0.130473
BLEU -3	0.0469856	0.056131
BLEU -4	0.022157	0.024011
METEOR	25.603077	**28.141276**
ROUGE-L	**29.575747**	**33.385576**
CIDEr	0.179349	0.224412

Table 5.3 Comparison of Individual scores generated using Inception V3 and VGG16

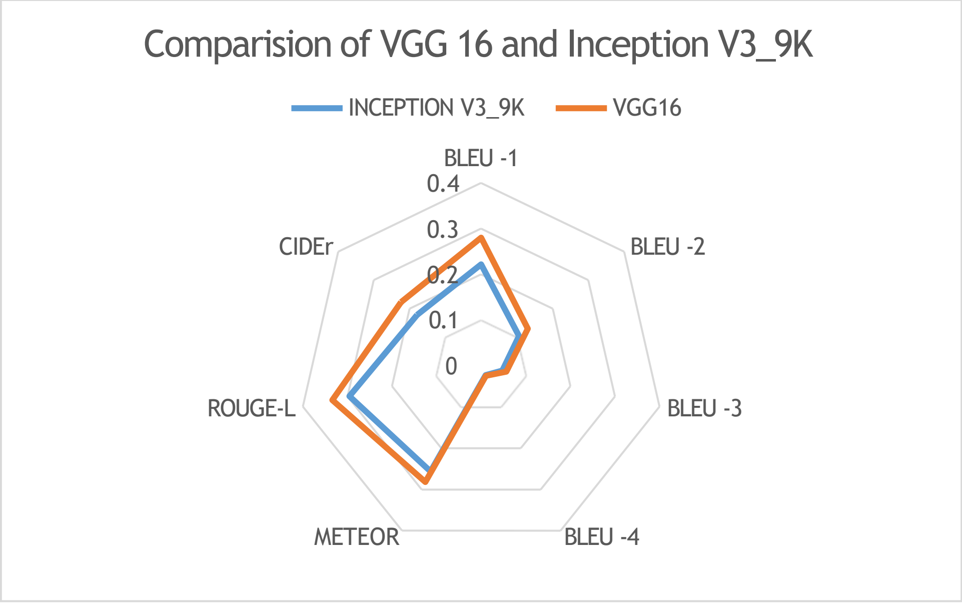
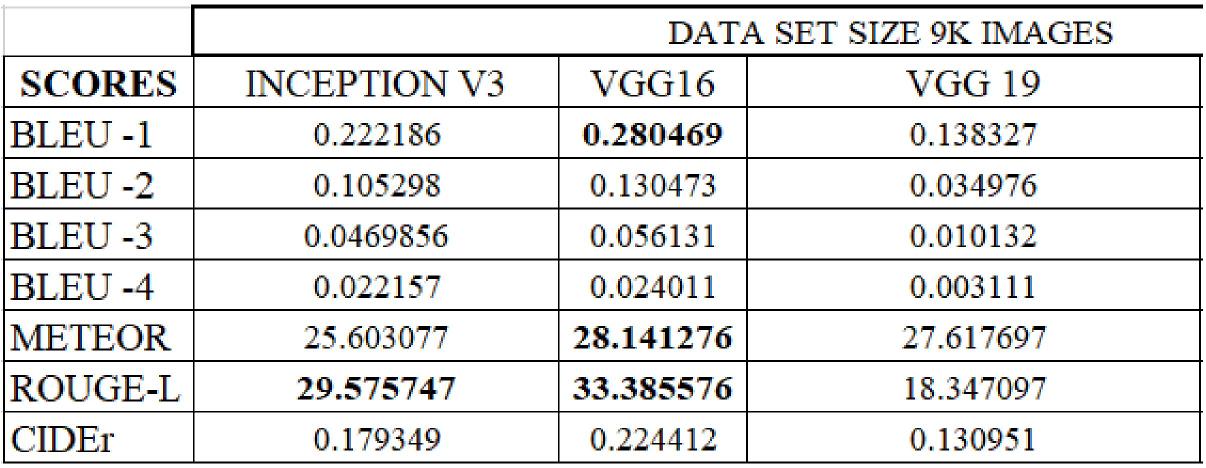


Figure 5.3 Comparison of VGG 16 and Inception V3 run on same data set sizes

The performance is better over all for Inception V3 when the data set size is increased 5 fold, figure 5.4 shows the overlap of BLUE metrics for different data set sizes for BLEU 3 and 4 for both algorithms. VGG 16 shows comparable performance on BLEU scores but the increased data set size makes Inception V3 the better performer. Table 5.4 shows individual values of both algorithms and different data set sizes.



	DATA SET SIZE 9K IMAGES		
SCORES	INCEPTION V3	VGG16	VGG 19
BLEU -1	0.222186	**0.280469**	0.138327
BLEU -2	0.105298	0.130473	0.034976
BLEU -3	0.0469856	0.056131	0.010132
BLEU -4	0.022157	0.024011	0.003111
METEOR	25.603077	**28.141276**	27.617697
ROUGE-L	**29.575747**	**33.385576**	18.347097
CIDEr	0.179349	0.224412	0.130951

Table 5.4 Comparison of Individual scores generated using InceptionV3, VGG16 and VGG19

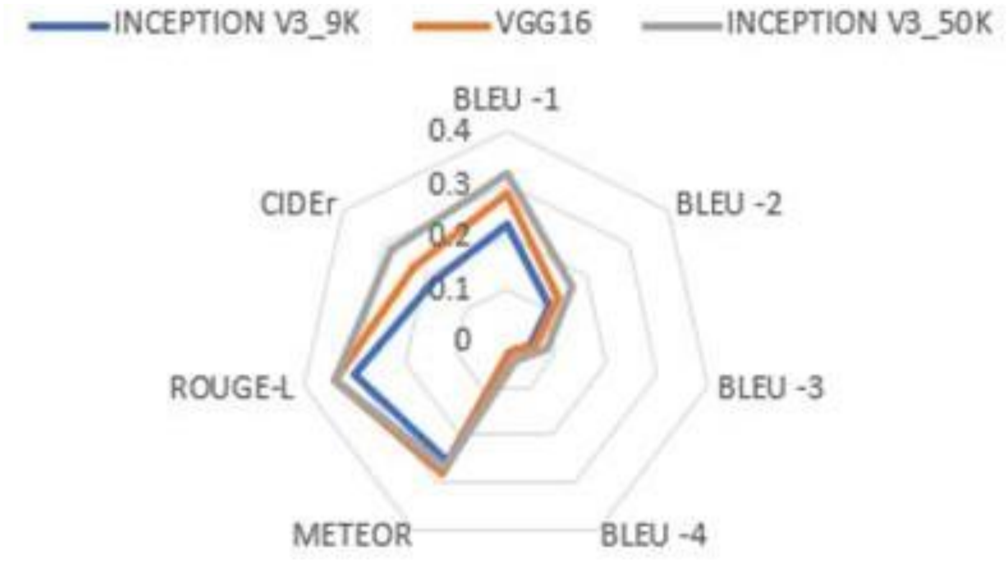
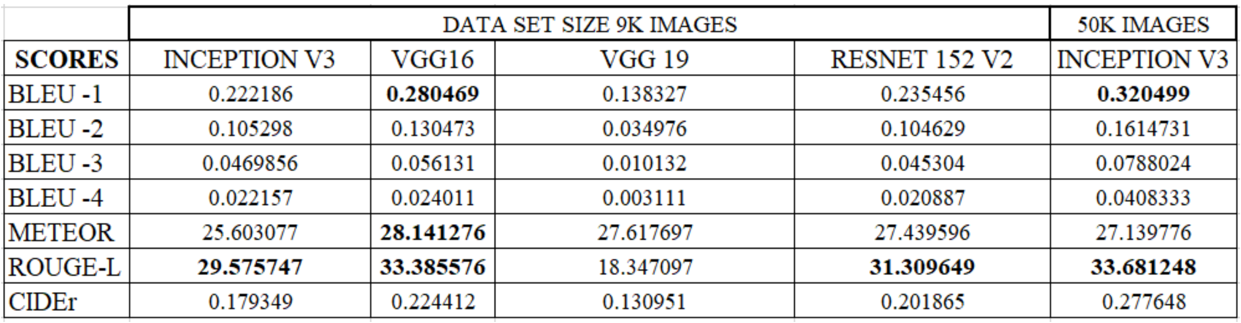


Figure 5.4 Comparison of Inception V3, VGG16 and VGG19 run on different data set sizes

5.2.4 *Comparison of algorithmic performance across metrics and data set sizes*

Observing the scores of all algorithms show VGG 16 to be the better performing for image captioning predictive capabilities and Inception V3 to give better performance when the data set size is significantly increased and makes it the next better performing algorithm for smaller image data set sizes. VGG 19 being a newer version of VGG 16 underperforms over its earlier version and also RESNET with comparable performance for METEOR. Table 5.5 highlights the scores where VGG 16 has shown significant performance on metrics and Inception V3 performing best over all other algorithm’s when the data sets size is increased by 500 percent.



	DATA SET SIZE 9K IMAGES				50K IMAGES
SCORES	INCEPTION V3	VGG16	VGG 19	RESNET 152 V2	INCEPTION V3
BLEU -1	0.222186	**0.280469**	0.138327	0.235456	**0.320499**
BLEU -2	0.105298	0.130473	0.034976	0.104629	0.1614731
BLEU -3	0.0469856	0.056131	0.010132	0.045304	0.0788024
BLEU -4	0.022157	0.024011	0.003111	0.020887	0.0408333
METEOR	25.603077	**28.141276**	27.617697	27.439596	27.139776
ROUGE-L	**29.575747**	**33.385576**	18.347097	**31.309649**	**33.681248**
CIDEr	0.179349	0.224412	0.130951	0.201865	0.277648

Table 5.5 Consolidated scores for algorithms used in the study and across data sizes

5.3 *Summary*

The scores generated involved various evaluation metrics that are most frequently used for image captioning studies which are BLEU scores with all its variants 1-4, METEOR, ROUGE- L and CIDE-r. The study was run for four algorithms Inception V3, VGG 16, VGG 19 and RESNET. The data set was consistent at 9K images for all algorithms. The limited resources and time availability was a hinderance to use a much larger data set size of 50K images to provide more training data sets for the algorithms. The larger data set size of 50K was run only for Inception V3 algorithms and no other algorithm during the course of the study. VGG 16 outperformed all other algorithms on all evaluation metrics and Inception V3 performed better over its previous score with an increase in the data set size. Inception V3 scores also performed better on all evaluation metrics with increased data set size compared with VGG 16, VGG 19 and RESNET.

6.1 *Conclusion and future work*

Inception V3 has performed best than all other algorithms due to increased data set size which allows more training data. Hence on all estimators Inception V3 has given the best performance. Meteor and Rouge scores have given a comparable result with the scores for the increased data set is only marginally better. BLEU scores have performed better than even an advanced algorithm like Inception V3, which wasn't anticipated for the same data set size. VGG 16 has given better performance than its newer version VGG 19 and the BLEU scores far outperform on all its variants 1-4, METEOR scores are comparable and ROUGE and CIDEr have also yielded better scores than VGG 19 . METEOR scores are comparable for all algorithms used and doesn't show much variance for an increase in the data set even by around 500%. VGG 16 has been the best performer for the 9K datasets over all other performance with increase in data set has given inception v3 only a marginal lead. METEOR and ROUGE scores are similar though there is a huge increase in the dataset size by almost 5 times, inception v3 has shown lesser performance though run on the same data set size which requires further understanding. It would be interesting to understand how the performance for VGG16 would be if the data set was increased would Inception V3 still come out better or VGG 16 though lesser advanced than Inception V3 would steal the show would make interesting research and could be attempted for future work.

REFERENCES

Anon (2021) *A Guide to ResNet, Inception v3, and SqueezeNet | Paperspace Blog*. [online] Available at: https://blog.paperspace.com/popular-deep-learning-architectures-resnet- inceptionv3-squeezenet/ [Accessed 26 Sep. 2021].

Anon (2021) *COCO - Common Objects in Context*. [online] Available at: https://cocodataset.org/#home [Accessed 27 Sep. 2021].

Anon (2021) *Google AI Blog: Improving Inception and Image Classification in TensorFlow*. [online] Available at: https://ai.googleblog.com/2016/08/improving-inception-and-image.html [Accessed 27 Sep. 2021].

Anon (2021) *Overcoming Challenges In Automated Image Captioning*. [online] Available at: https://www.ibm.com/blogs/research/2019/06/image-captioning/ [Accessed 9 Mar. 2021].

Banerjee, A., Krishan, R., Oyelade, O.N., El, A. and Ezugwu, S., (n.d.) Application of a Novel and Improved VGG-19 Network in the Detection of Workers Wearing Masks Recent citations A State-of-the-Art Survey on Deep Learning Methods for Detection of Architectural Distortion From Digital Mammography. *Journal of Physics: Conference Series PAPER • OPEN ACCESS*.

Bernardi, R., Cakici, R., Elliott, D., Erdem, A., Erdem, E., Ikizler-Cinbis, N., Keller, F., Muscat,
A. and Plank, B., (2016) Automatic description generation from images: A survey of models, datasets, and evaluation measures. *Journal of Artificial Intelligence Research*, 55, pp.409–442.

Biswas, R., Michael Barz, · and Sonntag, · Daniel, (n.d.) Towards Explanatory Interactive Image Captioning Using Top-Down and Bottom-Up Features, Beam Search and Re-ranking. *KI - Künstliche Intelligenz*, [online] 34, p.3. Available at: https://doi.org/10.1007/s13218-020-00679- 2.
Cascianelli, S., Costante, G., Ciarfuglia, T.A., Valigi, P. and Fravolini, M.L., (2018) Full-GRU Natural language video description for service robotics applications. *IEEE Robotics and Automation Letters*, 32, pp.841–848.

Du, J., Qin, Y., Lu, H. and Zhang, Y., (2018) Attend more times for image captioning. *arXiv*. Farhadi, A., Hejrati, M., Sadeghi, M.A., Young, P., Rashtchian, C., Hockenmaier, J. and Forsyth, D., (n.d.) *Every Picture Tells a Story: Generating Sentences from Images*.

Feng, Y., Ma, L., Liu, W. and Luo, J., (n.d.) *Unsupervised Image Captioning*.

Gong, Y., Wang, L., Hodosh, M., Hockenmaier, J. and Lazebnik, S., (n.d.) *Improving Image- Sentence Embeddings Using Large Weakly Annotated Photo Collections*.

Harold Li, L., Yatskar, M., Yin, D., Hsieh, C.-J. and Chang, K.-W., (n.d.) *Work in Progress VISUALBERT: A SIMPLE AND PERFORMANT BASELINE FOR VISION AND LANGUAGE*.

He, K., Zhang, X., Ren, S. and Sun, J., (2016) Deep residual learning for image recognition. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2016-Decem, pp.770–778.

He, K., Zhang, X., Ren, S. and Sun, J., (n.d.) *Deep Residual Learning for Image Recognition*.

He, K., Zhang, X., Ren, S. and Sun, J., (n.d.) *Deep Residual Learning for Image Recognition*. [online] Available at: http://image-net.org/challenges/LSVRC/2015/.

J. Curran, S. Clark, J.B., (n.d.) Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions. pp.33–36.

Karpathy, A. and Fei-Fei, L., (2017) Deep Visual-Semantic Alignments for Generating Image Descriptions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 394, pp.664– 676.

Khurana, K. and Mundada, S., (2018) Image Caption Generation : A Survey. 8.

Kiros, R., Salakhutdinov, R. and Zemel, R., (n.d.) *Multimodal Neural Language Models*.

Kiros, R., Salakhutdinov, R. and Zemel, R.S., (2014) Unifying Visual-Semantic Embeddings with Multimodal Neural Language Models.

Kumar, A. and Goel, S., (2017) A survey of evolution of image captioning techniques. *International Journal of Hybrid Intelligent Systems*, 143, pp.123–139.

Li, X., Yin, X., Li, C., Zhang, P., Hu, X., Zhang, L., Wang, L., Hu, H., Dong, L., Wei, F., Choi, Y. and Gao, J., (n.d.) *Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks*. Lin, D., (n.d.) Proceedings of the Fifteenth International Conference on Machine Learning. pp.296–304.

Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P. and Zitnick, C.L., (2014) Microsoft COCO: Common objects in context. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 8693 LNCSPART 5, pp.740–755.

M. Banko, V. Mittal, and M.W., (2000) Headline Generation Based on Statistical Translation. In: *38th Ann. Meeting Assoc. for Computational Linguistics*. pp.318–325.

Mao, J., Xu, W., Yang, Y., Wang, J., Huang, Z., Research, B. and Yuille, A., (n.d.) *DEEP CAPTIONING WITH MULTIMODAL RECURRENT NEURAL NETWORKS (M-RNN)*.
v. Ordonez, G. Kulkarni, T.L.B., (2011) Describing images using 1 million captioned photographs. *Advances in Neural Information Processing Systems*, Im2text, pp.1143–1151. Ordonez, V., Kulkarni, G. and Berg, T.L., (n.d.) *Im2Text: Describing Images Using 1 Million Captioned Photographs*.

Plummer, B.A., Wang, L., Cervantes, C.M., Caicedo, J.C., Hockenmaier, J. and Lazebnik, S., (n.d.) *Flickr30k Entities: Collecting Region-to-Phrase Correspondences for Richer Image-to- Sentence Models*.

Ren, L. and Hua, K., (2018) Improved Image Description Via Embedded Object Structure Graph and Semantic Feature Matching. In: *2018 IEEE International Symposium on Multimedia (ISM)*. pp.73–80.

Ren, L., Qi, G. and Hua, K., (2019) Improving Diversity of Image Captioning Through Variational Autoencoders and Adversarial Learning. In: *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*. pp.263–272.

Shi, Z. and Zou, Z., (2017) Can a Machine Generate Humanlike Language Descriptions for a Remote Sensing Image? *IEEE Transactions on Geoscience and Remote Sensing*, 556, pp.3623–3634.

Srinivasan, L. and Sreekanthan, D., (2018) *Image Captioning-A Deep Learning Approach*. [online] *International Journal of Applied Engineering Research*, Available at: http://www.ripublication.com [Accessed 9 Mar. 2021].

Su, W., Zhu, X., Cao, Y., Li, B., Lu, L., Wei, F. and Dai, J., (n.d.) *VL-BERT: PRE-TRAINING OF GENERIC VISUAL-LINGUISTIC REPRESENTATIONS*.
Sun, C., Gan, C. and Nevatia, R., (n.d.) *Automatic Concept Discovery from Parallel Text and Visual Corpora*.

Suresh, R. and Keshava, N., (2019) A Survey of Popular Image and Text analysis Techniques. *CSITSS 2019 - 2019 4th International Conference on Computational Systems and Information Technology for Sustainable Solution, Proceedings*.

Szegedy, C., Vanhoucke, V., Ioffe, S. and Shlens, J., (n.d.) *Rethinking the Inception Architecture for Computer Vision*.

Vinyals Google, O., Toshev Google, A., Bengio Google, S. and Erhan Google, D., (n.d.) *Show and Tell: A Neural Image Caption Generator*.

Waghmare, P. and Shinde, S., (2020) *International Conference on Communication and Information Processing Artificial Intelligence Based Image Caption Generation*. [online] Available at: https://ssrn.com/abstract=3648847 [Accessed 9 Mar. 2021].

Wang, H., Zhang, Y. and Yu, X., (2020) An Overview of Image Caption Generation Methods. [online] Available at: https://doi.org/10.1155/2020/3062706.

Xu, K., Ba, J.L., Kiros, R., Cho, K., Courville, A., Salakhutdinov, R., Zemel, R.S. and Bengio, Y., (2015) Show, attend and tell: Neural image caption generation with visual attention. *32nd International Conference on Machine Learning, ICML 2015*, 3, pp.2048–2057.

Xu, K., Ba, J.L., Kiros, R., Cho, K., Courville, A., Salakhutdinov, R., Zemel, R.S. and Bengio, Y., (n.d.) *Show, Attend and Tell: Neural Image Caption Generation with Visual Attention*.

Yan, S., Xie, Y., Wu, F., Smith, J.S., Lu, W. and Zhang, B., (2018) Image captioning via a hierarchical attention mechanism and policy gradient optimization. *arXiv*, 148, pp.1–13.

Yan, S., Xie, Y., Wu, F., Smith, J.S., Lu, W. and Zhang, B., (n.d.) *Image Captioning via a Hierarchical Attention Mechanism and Policy Gradient Optimization*.

Zakir Hossain, M.D., Sohel, F., Shiratuddin, M.F. and Laga, H., (2018) A comprehensive survey of deep learning for image captioning. *arXiv*, 00.