**Department of Computer Science**

**BSc (Hons) Business Computing (with Option if appropriate)**

**BSc (Hons) Computer Science (with Option if appropriate)**

Academic Year 2019 - 2020

The concise and insightful title for the project to which this dissertation pertains

Student Name

01234567

A report submitted in partial fulfilment of the requirements for the degree of

Bachelor of Science

# Abstract

Image caption generation has been a difficult computer science downside, for a considerable amount within the 1900's wherever a matter description should be generated for a given image. it's a synthetic Intelligence downside desegregation varied fields of AI as well as language process (NLP) that aids in understanding the idea of the several image and laptop Vision that gives with deep learning models to coach on our several datasets to attain the specified results.

Image captioning is a crucial downside of computer science that not solely aims to assist computers gain human like intelligence however conjointly helps in providing technology-based resolution for higher reading, teaching and a step shut for visually impaired, deaf individuals.

In the recent years, deep learning is rising because the new upgrade of AI that’s provides with the flexibleness of employing a single end-to-end model are often created or retrained on a brand new information set to assist in predicting captions of latest pictures while not the necessity for superfluous and sophisticated data preparation or model creation techniques of the past.

Deep learning has appointed best degree of results for our downside. Our project is predicated on Flickr 8k dataset, that gives with eight thousand pictures to coach our deep learning models for captioning. This dataset consists of pictures from all aspects of nature, as well as each living and dead aspects of life. This project can demonstrate the marvels of deep learning exploitation python and boa within the job of Image captioning.

In the end, we are going to be providing an outline, conclusion and analysis of our deep learning model supported the results that are obtained by providing our model with unknown pictures.

This analysis can give with a definitive direction towards our initial call of selecting deep learning to resolve the given downside.

# Acknowledgements

First and foremost, praises and due to the God, the Almighty, for His showers of blessings throughout my analysis work to finish the analysis with success.

I would prefer to specific my deep and sincere feeling to my analysis supervisor, Dr. Fang Wang, my professor at Brunel University London. Emma Batley, Simon Kent, Rob Macredie, Alina Miron for giving me the chance to try to do analysis and providing valuable steerage throughout this analysis. His dynamism, vision, sincerity and motivation have deeply galvanized me and my fellow classmates. He has instructed our class, the methodology to hold out the analysis and to gift the analysis works as clearly as attainable. it had been a good privilege and honour to figure and study beneath his steerage. I’m extraordinarily grateful for what he has offered to us and many more successful batches of students before us. I’d conjointly prefer to convey him for his relationship, empathy, and nice sense of humour. I’m extending my dear due to his partner, family for his or her acceptance and patience throughout the discussion I had with him on analysis work and thesis preparation.

I am extraordinarily grateful to my folks for his or her love, prayers, caring and sacrifices for educating and making me ready for my future. Also, I specific my due to my sisters, brother, sister in law and brother in laws for his or her support and valuable prayers.

I would prefer to say due to my friends and analysis colleagues, for his or her constant encouragement. I specific my special thanks to my best friend, for his real support throughout this analysis work.

I am extending my due to the BSc. students of Brunel University London, for their support throughout my analysis work. I conjointly convey all the employees of analysis section of Brunel University London, Tirunelveli for his or her kindness.

I convey the management of C.S. department for their support to try to this work.

Finally, my thanks head to all the those that have supported me to finish the analysis work directly or indirectly.

I certify that the work presented in the dissertation is my own unless referenced.

Signature \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Date \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Total Words: 2005

# Table of Contents

Abstract i

Acknowledgements ii

Table of Contents iv

List of Tables v

List of Figures vi

1 Introduction 1

1.1 Aims and Objectives 1

1.1 Project Approach 4

1.2 Dissertation Outline 6

2 Background 7

2.1 Tables 7

2.2 Figures 8

3 Methodology (Or Approach) 9

4 Design ( Or What you did Part One) 9

5 Implementation (Or What you did Part Two) 12

6 Testing and Evaluation 20

7 Conclusions 20

7.1 Future Work 23

References 25

Appendix A Personal Reflection 26

A.1 Reflection on Project 26

A.2 Personal Reflection 26

Appendix B Appendices 27

B.1 More relevant material 27

# List of Tables

Table 1 - A risk analysis table

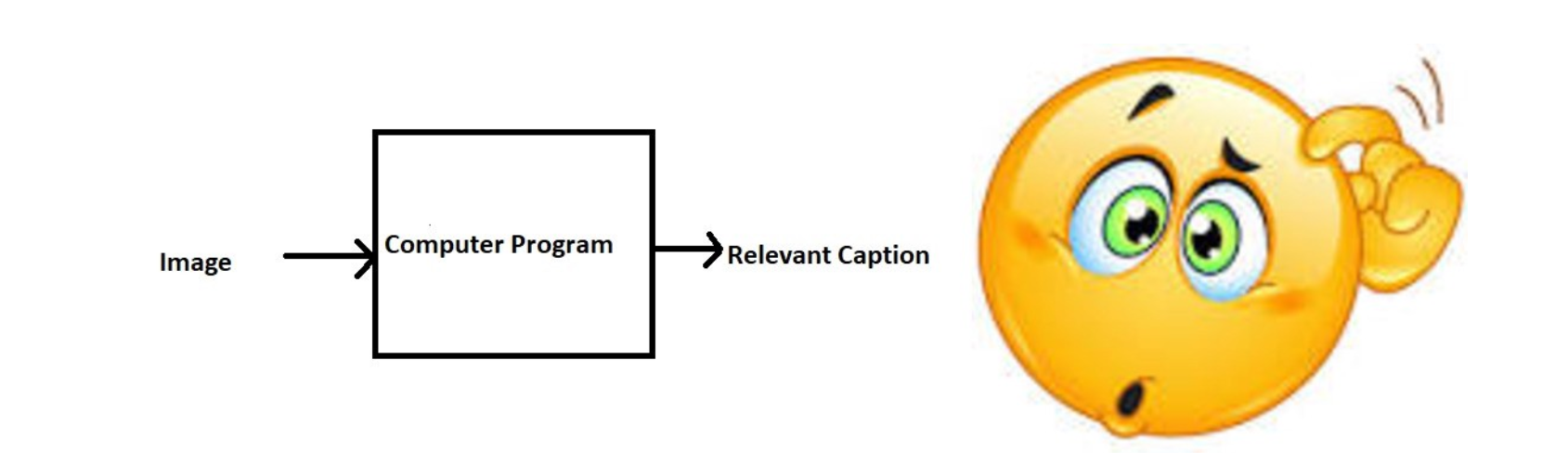
# List of Figures

Figure 1 - A generic model of the Final Year Project 5

# Introduction

Computer scientists and researchers square measure engaged on 'describing images' for quite a terribly very long time, throughout their initial experimentation that they started from object detection, they completed that associate rule offers( that gives, that has) with simply the names or labels for the objects gift at intervals the image does not give with a person's like description and thus, it does not provides a smart impression.

In the previous few years, 2 keywords have gotten the vocabulary of just about each article written underneath the image captioning topic—novel and linguistics. These keywords square measure vital for determination the most important challenge during this exercise i.e. generating a caption in a very manner that it'd be indivisible from human written ones. linguistics implementation is meant to style a clean manner of injecting sentiment into this image captioning system. Novel objects should be enclosed for the enlargement of eventualities. There are many insights on why this can be still associate degree open issue. initial of all, typically models square measure designed on terribly knowledge sets, that don't cowl all doable eventualities and aren't applicable in describing various atmosphere. a similar with vocabulary because it incorporates a restricted range of words and their combos. Second, models square measure typically thought to perform on one terribly task, whereas humans square measure ready to work on several tasks at the same time. Many Scientists within the past have already tried to beat this drawback and has provided an answer though it had been not additional continued. Another nice approach for addressing unseen knowledge, because it is presently not possible to feed all existing knowledge into the machine, was projected within the recent years. long learning relies on a questioning approach i.e. creating a discussion directly with the user or within the model. This approach depends on a natural manner of human communication; from time of life youngsters largely learn by asking queries. The model is meant to find out conjointly sort of a child—by asking terribly queries and learning from the answers. This technique falls underneath the question respondent topic—a literature analysis thorough can be done on this subject as here we've got conferred solely what appeared throughout this study on image captioning. this may be targeted as a separate drawback; however, it conjointly makes a good impact in image captioning.



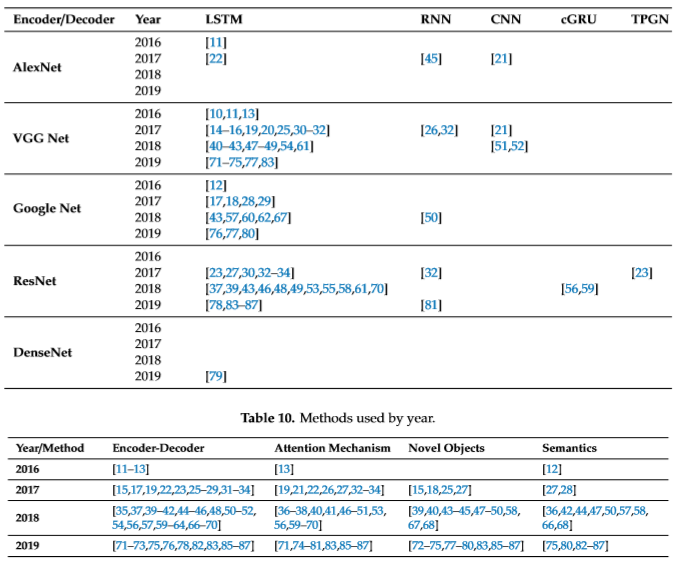
The main aim of language or extra considerably applied science is to make machines assume, speak and behave at intervals the way like humans. that's why image descriptions (Natural language) remained a challenge to be resolved. Scientists try a selection of techniques similarly as neural nets, NLP, Machine learning etc. since early 2014 in step with a probe paper (by Raimonda Staniute and Dmitrij Sesok).

However, the interest in image captioning died down throughout the late 2014 till 2015 but it rose in 2018 and onward and it's still a potential challenge as new information and data keeps bring together throughout this ever-quick paced international village. The scope, profits and benefits of intelligent machines and anthropomorphic technology is predominating and plain, be it for serving to disabled youngsters with education and learning, factories, healthcare, farming, virtual assistants for recent aged folks etc.

Even though, this precise field of school is rising nevertheless not many of us have grasped the opportunities to analysis and be ingenious within the scope of intelligent machines as they must have. Machines of these days, still lack in their natural humanly manner of communication and no machine has absolutely passed the mathematician, Alan Mathison Turing, mathematician} check until date ("A Turing check may be a technique or technique of inquiry in AI (AI) to crucial whether or not or not a pc is capable of thinking sort of a soul. The check is known as when Alan Turing, the founding father of the Turning check associated an English computer user, decipherer, scientist and theoretical scientist.").

With the event of deep learning, the mix of pc vision and linguistic communication method has aroused nice attention within the past few years. Image captioning may be a representative of this filed, that makes the pc learn to use one or additional sentences to know the visual content of a picture. The important description generation method of high-level image linguistics needs not solely the popularity of the article and also the scene, however the power of analyzing the state, the attributes and also the relationship among these objects. though image captioning may be a difficult and tough task; a great deal of researchers has achieved important enhancements.

Description of pictures may be a key drawback and an enquiry issue in image captioning.



## Aims and Objectives

The aim of this project is to provide a deep learning-based Image captioning model that has been trained on majority of object classes provided by open source datasets.

Whereas the objectives include:

* Data Collection: There are many open source datasets available for this problem, like Flickr 8k (containing8k images), Flickr 30k (containing 30k images), MS COCO (containing 180k images), etc.

But for the purpose of this case study, I have used the Flickr 8k dataset. This dataset contains 8000 images each with 5 captions (as we have already seen in the Introduction section that an image can have multiple captions, all being relevant simultaneously).

These images are bifurcated as follows:

* + - Training Set — 6000 images
    - Dev Set — 1000 images
    - Test Set — 1000 images
* Data Processing: By the term data processing we mean that we will be first separating the dataset into training and testing sets after which the data will then be prepared in the appropriate format to be provided for the training of deep learning model.
* Model training: The proposed deep learning model will then be trained
* Model Testing:
* Model Evaluation:

## Project Approach

Given a brand-new image, a picture captioning formula ought to output an outline regarding this image at a linguistics level. For the image captioning task, humans will simply perceive the image content and categorical it within the variety of linguistic communication sentences in step with specific needs; but, for computers, it needs the integrated use of image process, pc vision, linguistic communication process and alternative major areas of analysis results. The challenge of image captioning is to style a model that may absolutely use image info to come up with additional human-like made image descriptions.

The important description generation method of high-level image linguistics needs not solely the understanding of objects or scene recognition within the image, however additionally the power to analyses their states, perceive the link among them and generate a semantically and syntactically correct sentence. it's presently unclear however the brain understands a picture and organizes the visual info into a caption. Image captioning involves a deep understanding of the globe and that things square measure salient components of the total.

If we tend to square measure ready to perform automatic image annotations, then this will have each sensible and theoretical advantages. within the current social development method, the foremost necessary factor is that the large information that exists on the net. Most of that information square measure totally different from ancient information, and media information occupies an outsized proportion. they're typically generated from net merchandise like social networks or print media. with the exception of the very fact that humans will directly method these media pictures, the helpful info that the machine will presently collect from them is proscribed and it's tough to help humans in more work. Image captioning tasks, if they're correct enough, will handle large amounts of media information and generate human linguistic communication descriptions that square measure additional acceptable to humans. The machine is going to be ready to higher assist citizenry to use these media information to try and do additional things.

Intelligent observance allows the machine to spot and verify the behavior of individuals or vehicles within the captured scene and generate alarms underneath acceptable conditions to prompt the user to react to emergencies and forestall excess accidents. for instance, in channel observance, it collects the golf green operations and criminal activities, monitors the conditions of the golf green, and promptly discovers the conditions of the waterway operations, traffic conditions, criminal sand mining, and also the use of navigation channels. Then report the case to the command center for programming and stop criminal activities in an exceedingly timely manner. Image captioning may be applied to the current side. Through the image captioning strategies, the machine will perceive the scenes it captures, so it will reply to specific things or apprise users in an exceedingly timely manner supported human settings.

The existing image captioning methodology supported deep learning or machine learning needs heaps of marked coaching samples. In sensible applications, it's needed that a text description of a plurality of languages will be provided for the image to fulfill the requirements of various language users. At present, there are several coaching samples represented in English and Chinese texts, however there are few mark-ups in alternative language text descriptions. If the matter description of every language within the image is dispensed, manual marking would require heaps of workforce and time. Therefore, the way to implement cross language text. A better understanding of the project is provided in the upcoming chapters specially the implementation part.

## Dissertation Outline

Chapter 2 discusses the background for my project and identifies some key techniques that can be adopted during the development of the proposed solution.

Chapter 3 explains how the project will be undertaken, it is focused towards the detailed view of methodology or approach of the project.

Chapter 4 begins with the detailed analysis of our projects design and all the components respectively.

Chapter 5 is the most important chapter with reference to the work done. This is the main chapter that informs the readers about our input and hard work in this current problem.

Chapter 6 shows us the results of testing and evaluation, where as chapter 7 concludes our project report and give a brief towards the future prospect of how this problem can be better improvised.

# Background

With the advancements of science and technology and also, they would like for the event of human life, robots are utilized in additional and additional industries. Auto-pilot robots will be showing intelligence avoid obstacles, amendment lanes and pedestrians supported the road conditions in step with the encircling driving surroundings they observe. additionally, to safe and economical driving, it's additionally potential to perform operations like automatic parking. Liberating the driver’s eyes and hands will greatly facilitate people’s lives and cut back safety accidents. If the machine needs to try and do the work higher, it should move with humans higher. The machine will tell humans what it sees, and humans then perform acceptable process supported machine feedback. to finish these tasks, we want to have faith in automatic generation of image descriptions.

When a user uploads an image, the image must be illustrated and annotated which might be simply found by the opposite users. the normal methodology is to retrieve the foremost similar image within the information for annotation, however this methodology typically ends up in incorrectly annotated pictures. Besides, video has currently become an imperative a part of people’s lives. so, as to fancy movies higher, several movies currently need subtitles. Every year, there ar an oversized range of videos made worldwide. These videos are composed of tens of thousands of images. Therefore, image and video annotation are a significant task. the automated generation of the image description will method all the video frames, so mechanically generate the corresponding text description in line with the content of the video frame, which might greatly cut back the work of the video employee and may complete the video annotation work with efficiency and effectively. additionally, image and video annotation may also facilitate visually impaired individuals to grasp an oversized range of videos and footage on the net. The image description is generated mechanically within the aspects of intelligent watching, human-computer interaction, image and video annotation. this is often solely a part of the image captioning applications. In short, image captioning will so be applied in several aspects of people’s lives, which might greatly improve labor potency and facilitate people’s life, production and learning.

A screenshot of a cell phone

Description automatically generated

# Methodology (Or Approach)

Deep learning strategies have incontestable progressive results on caption generation issues. what's most spectacular regarding these strategies may be a single end-to-end model may be outlined to predict a caption, given a photograph, rather than requiring subtle information preparation or a pipeline of specifically designed models. There are a unit several models to settle on from. during this case, we'll use the Oxford Visual pure mathematics cluster, or VGG, model that won the ImageNet competition in 2014. we'll use a pre-trained model to interpret the content of the photos. Convolutional networks (ConvNets) presently set the state of the art in visual recognition. The aim of this project is to research however the ConvNet depth affects their accuracy within the large-scale image recognition setting.

The issue of image caption generation involves producing a readable, clear and apothegmatic description of the subject matter of a picture. It is a difficult computer science drawback because it needs each technique from laptop vision to interpret the contents of the photograph and techniques from tongue process to get the matter description. Recently, deep learning ways have achieved progressive results on this difficult drawback. The results are therefore spectacular that this drawback has become a typical demonstration drawback for the capabilities of deep learning.

The main contribution by Oxford professors and Machine Learning scientists WHO worked on these models may be a rigorous analysis of networks of accelerating depth, that shows that a big improvement on the prior-art configurations may be achieved by increasing the depth to 16-19 weight layers, that is well deeper than what has been employed in the previous art. to cut back the amount of parameters in such terribly deep networks, we have a tendency to use terribly tiny 3×3 filters all told convolutional layers.

VGG16 may be a convolutional neural network model planned by K. Simony an and A. Zisserman from the University of Oxford within the paper “Very Deep Convolutional Networks for Large-Scale Image Recognition”. The model achieves ninety-two-point seven percent, top-5 check accuracy in ImageNet, that may be a dataset of over fourteen million pictures happiness to a thousand categories. it absolutely was one in all the known model submitted to ILSVRC-2014. It makes the development over Alex Net by replacement giant kernel-sized filters (11 and five within the initial and second convolutional layer, respectively) with multiple 3×3 kernel-sized filters one once another. VGG16 was trained for weeks and was victimisation NVIDIA Titan Black GPU’s.

ImageNet may be a dataset of over fifteen million tagged high-resolution pictures happiness to roughly twenty-two thousand classes. the photographs were collected from the online and tagged by human labelers victimization Amazon’s Mechanical Turk crowd-sourcing tool. beginning in 2010, as a part of the Pascal Visual Object Challenge, associate degree annual competition known as the ImageNet Large-Scale Visual Recognition Challenge (ILSVRC) has been command. ILSVRC uses a set of ImageNet with roughly a thousand pictures in every of a thousand classes. In all, there are a unit roughly one point two million coaching pictures, 50,000 validation pictures, and 150,000 testing pictures. ImageNet consists of variable-resolution pictures. Therefore, the photographs are down sampled to a hard and fast resolution of 256×256. Given an oblong image, the image is rescaled and cropped out the central 256×256 patch from the ensuing image.

# Design (Or What you did Part One)

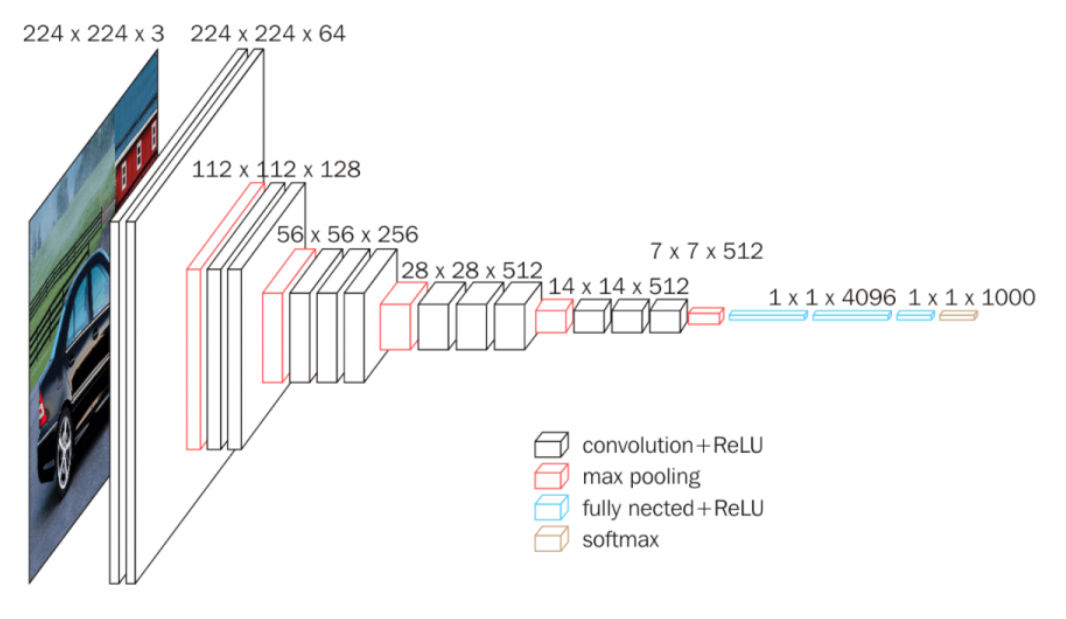
The architecture of the deep learning model is as follows: The input to cov1 layer is of fastened size 224 x 224 RGB image. The image is well-versed a stack of convolutional (conv.) layers, wherever the filters were used with a really little receptive field: 3×3 (which is that the smallest size to capture the notion of left/right, up/down, center). In one in all the configurations, it conjointly utilizes 1×1 convolution filter, which may be seen as a linear transformation of the input channels (followed by non-linearity). The convolution stride is fastened to one pixel; the spatial cushioning of conv. layer input is such the spatial resolution is preserved when convolution, i.e. the cushioning is 1-pixel for 3×3 conv. layers. spatial pooling is allotted by 5 max-pooling layers, that follow a number of the conv. layers (not all the conv. layers are followed by max-pooling). Max-pooling is performed over a 2×2 component window, with stride two.

Three Fully Connected (FC) layers follow a stack of convolutional layers (which features a totally different depth in several architectures): the primary 2 have 4096 channels every, the third performs one thousand-way ILSVRC classification and so contains 1000 channels (one for every class). the ultimate layer is that the soft-max layer. The configuration of the totally connected layers is that the same altogether networks.

All hidden layers are equipped with the rectification (ReLU) non-linearity. it's conjointly noted that none of the networks (except for one) contain native Response social control (LRN), such standardization doesn't improve the performance on the ILSVRC dataset, however results in redoubled memory consumption and computation time.

VGG16 considerably outperforms the previous generation of models within the ILSVRC-2012 and ILSVRC-2013 competitions. The VGG16 result's conjointly competitive for the classification task winner (Google Net with half-dozen.7% error) and well outperforms the ILSVRC-2013 winning submission Clarified, that achieved eleven point two% with external coaching knowledge and eleven.7% while not it. regarding the single-net performance, VGG16 design achieves the simplest result (7.0% look at error), outperforming one Google Net by zero.9%.

It was incontestable that the illustration depth is helpful for the classification accuracy, which progressive performance on the ImageNet challenge dataset are often achieved employing a standard Conv Net design with well redoubled depth.



# Implementation (Or What you did Part Two)

Keras provides this pre-trained model directly. Note, the primary time you employ this model, Keras can transfer the model weights from the web, that are regarding five hundred Megabytes, this could take some minutes betting on your net affiliation.

We may use this model as a part of a broader image caption model. the matter is, it's an outsized model and running every picture through the network each time we wish to check a replacement language model configuration (downstream) is redundant. Instead, we are able to pre-compute the “photo features” mistreatment the pre-trained model and save them to file. we are able to then load these options later and feed them into our model because the interpretation of a given picture within the dataset. it's no completely different to running the picture through the total VGG model; it's simply we are going to have done it once ahead. This is associate degree optimization which will create coaching our models quicker and consume less memory.

A close up of text on a white background

Description automatically generated

The implementation of the algorithm is divided in a few steps, the first begins with preparing the data for the model. We can load the VGG model in Keras using the VGG category. We are going to take away the last layer from the loaded model, as this is often the model want to predict a classification for a photograph. We tend to aren't fascinated by classifying pictures, however we tend to have an interest within the content of the picture right before a classification is created. These are the “features” that the model has extracted from the picture. Keras additionally provides tools for reshaping the loaded picture into the popular size for the model.

The first perform to use is called extract\_features () that, provided a directory name, can load every picture, prepare it for VGG, and collect the anticipated options from the VGG model. The image options are a one-dimensional, four thousand ninety-six component vector. The perform returns a lexicon of image symbol to image options. we are going to be job this perform to arrange the picture information for testing our model, then save the ensuing lexicon to a file named ‘features.pkl ‘. Running this information preparation step could take a minute counting on your hardware, maybe one hour on the electronic equipment with a contemporary digital computer. At the tip of the run, you'll have the extracted options hold on in ‘features.pkl ‘for later use. This file is concerning 127 Megabytes in size.

Next we'll continue with making ready the matter information. The dataset contains multiple descriptions for every photograph and therefore the text of the descriptions needs some borderline clean-up. First, we'll load the file containing all of the descriptions. Each image incorporates a distinctive symbol. This symbol is employed on the image file name and within the document of descriptions. Next, we'll step through the list of image descriptions. Below defines a operate load\_descriptions() that, given the loaded document text, can come a wordbook of image identifiers to descriptions. every image symbol maps to an inventory of 1 or additional matter descriptions.

Next, we'd like to wash the outline text. The descriptions area unit already tokenized and simple to figure with. We will clean the text within the following ways in which so as to scale back the dimensions of the vocabulary of words we'll got to work with: Convert all words to minuscular. Remove all punctuation. Remove all words that area unit one character or less long (e.g. ‘a’). Remove all words with numbers in them. We will be process the clean\_descriptions() operate at this time, given the wordbook of image identifiers to descriptions, steps through every description and cleans the text.

Once clean, we will summarize the dimensions of the vocabulary. Ideally, we would like a vocabulary that's each communicative and as little as attainable. A smaller vocabulary can lead to a smaller model which will train quicker. For reference, we will rework the clean descriptions into a group and print its size to urge a plan of the dimensions of our dataset vocabulary. Finally, we will save the wordbook of image identifiers and descriptions to a brand-new file named descriptions.txt, with one image symbol and outline per line.

Here we'll outline the save\_descriptions () operate that, given a wordbook containing the mapping of identifiers to descriptions and a file name, saves the mapping to file. Putting this all at once, the entire listing is provided below. **(PHOTOES)** Running the instance 1st prints the quantity of loaded image descriptions (8,092) and therefore the size of the clean vocabulary (8,763 words). Finally, the clean descriptions area unit written to ‘descriptions.txt ‘.

Taking a glance at the file, {we can, we will, we area unit able to} see that the descriptions are prepared for modelling. The order of descriptions in your file could vary. The next step is to develop the deep learning model, we will outline the deep learning model and work it on the coaching dataset. This step is split into the subsequent parts: -Loading information. -Defining the Model. -Fitting the Model. -Testing the Model. First, we tend to should load the ready image and text information so we will use it to suit the model. We area unit about to train the info on all of the photos and captions within the coaching dataset. whereas coaching, we tend to area unit about to monitor the performance of the model on the event dataset and use that performance to make your mind up once to save lots of models to file. The train and development dataset are predefined within the Flickr\_8k.trainImages.txt and Flickr\_8k.devImages.txt files severally, that each contain lists of image file names. From these file names, we will extract the image identifiers and use these identifiers to filter photos and descriptions for every set.

The operate load\_set () below can load a pre-defined set of identifiers given the train or development sets file name. Now, we will load the photos and descriptions mistreatment the pre-defined set of train or development identifiers. This is wherever the operate load\_clean\_descriptions (), are dead within the code that hundreds the clean text descriptions from ‘descriptions.txt ‘for a given set of identifiers and returns a wordbook of identifiers to lists of text descriptions. The model (we can, we'll, we are going to) develop will generate a caption given a photograph, and therefore the caption is generated one word at a time.

The sequence of antecedent generated words is provided as input. Therefore, we'll want a ‘first word’ to kick-off the generation method and a ‘last word ‘to signal the tip of the caption. We will be using two strings for this purpose namely, ‘start seq ‘and ‘end seq’. These tokens area unit additional to the loaded descriptions as they're loaded. it's vital to try and do this currently before we tend to write in code the text, so the tokens also are encoded properly.

Next, we will load the image options for a given dataset. Below defines a operate named load\_photo\_features () that hundreds the whole set of image descriptions, then returns the set of interest for a given set of image identifiers. This is not terribly efficient; withal, this may get USA up and running quickly. We can pause here and check everything developed to date. Running this instance 1st hundreds the half dozen thousand image identifiers within the check dataset. These options area unit then accustomed filter and cargo the clean description text and therefore the pre-computed image options. The description text can get to be encoded to numbers before it will be bestowed to the model as in input or compared to the model’s predictions.

The first step in cryptography the info is to form a uniform mapping from words to distinctive number values. Keras provides the Tokenizer category which will learn this mapping from the loaded description information. Here we'll outline the to\_lines () to convert the wordbook of descriptions into an inventory of strings and therefore the create\_tokenizer () operate which will work a Tokenizer given the loaded image description text. We can currently write in code the text. Each description is split into words. The model is provided one word and therefore the image and generate subsequent word. Then the primary 2 words of the outline are provided to the model as input with the image to get subsequent word.

This is however the model are going to be trained. For example, the input sequence “little woman running in field” would be split into half-dozen input-output pairs to coach the model.

Later, once the model is utilized to get the caption descriptions, the brought forth words are going to be added and recursively allowed for as input to get a caption for an image or photograph. The operate below named create\_sequences (), given the tokenizer, a most sequence length, and therefore the wordbook of all descriptions and photos, can remodel the information into input-output pairs of information for coaching the model. There are a unit 2 input arrays to the model: one for pic options and one for the encoded text. there's one output for the model that is that the encoded next word within the text sequence. The input text is encoded as integers, which is able to be fed to a word embedding layer. The picture options are going to be fed on to another a part of the model. The model can output a prediction, which is able to be a chance distribution over all words within the vocabulary.

The output knowledge can so be a one-hot encoded version of every word, representing Associate in Nursing perfect chance distribution with zero values in the slightest degree word positions except the particular word position, that includes a worth of one. We will have to be compelled to calculate the utmost variety of words within the longest description. a brief helper operates named max\_length () is outlined below. We currently have enough to load {the knowledge, the info, the information} for the coaching and development knowledge sets and remodel the loaded data into input-output pairs for fitting a deep learning model.

Defining the Model: We will outline a deep learning supported the “merge-model” delineated by brandy Tanti, et al. in their 2017 papers: -Where to place the Image in a picture Caption Generator, 2017. -What is that the role of Recurrent Neural Networks (RNNs) in a picture Caption Generator? 2017.

A brief summary of their analysis is given as follows: The encoder-decoder perennial neural spec has been shown to be effective at the caption generation drawback. The implementation of this design is often distilled into inject and merge based mostly models, and each build totally different assumptions concerning the role of the perennial neural network in addressing the matter. A standard encoder-decoder perennial neural spec is employed to deal with the image caption generation drawback.

This involves 2 elements:

* Encoder: A network model that reads the photograph input and encodes the content into a fixed-length vector mistreatment an interior illustration.
* Decoder: A network model that reads the encoded photograph and generates the matter description output.

Generally, a convolutional neural network is employed to cypher the photographs and a perennial neural network, like an extended immediate memory network, is employed to either cypher the text sequence generated thus far, and /or generate subsequent word within the sequence. There are a unit many ways to understand this design for the matter of caption generation. It is common to use a pre-trained convolutional neural network model trained on a difficult photograph classification drawback to cypher the photograph.

The pre-trained model is often loaded, the output of the model removed, and therefore the cognitive content of the photograph used because the cryptography or cognitive content of the input image. It is additionally common to border the matter specified the model generates one word of the output matter description, given each the photograph and therefore the description generated thus far as input. during this framing, the model is termed recursively till the whole output sequence is generated. This framing is often enforced mistreatment one in every of 2 architectures, known as by brandy Tanti, et al. because the inject and therefore the merge models.

* Inject Model: The inject model combines the encoded kind of the image with every word from the text description generated so-far. The approach uses the perennial neural network as a text generation model that uses a sequence of each image and word data as input so as to get subsequent word within the sequence. In these ‘inject’ architectures, the image vector (usually derived from the activation values of a hidden layer in an exceedingly convolutional neural network) is injected into the RNN, for instance by treating the image vector on a par with a ‘word’ and as well as it as a part of the caption prefix. This model combines the considerations of the image with every input word, requiring the encoder to develop Associate in Nursing cryptography that includes each visual and linguistic data along. In Associate in Nursing inject model, the RNN is trained to predict sequences supported histories consisting of each linguistic and sensory activity options. Hence, during this model, the RNN is primarily to blame for image-conditioned language generation.

A screenshot of a cell phone

Description automatically generated

* Merge Model: The merge model combines each the encoded kind of the image input with the encoded kind of the text description generated thus far. The combination of those 2 encoded inputs is then employed by an awfully straightforward decoder model to get subsequent word within the sequence. The approach uses the perennial neural network solely to cypher the text generated thus far. In merge design, the image is omitted of the RNN subnetwork, specified the RNN handles solely the caption prefix, that is, handles solely strictly linguistic data. once the prefix has been vectorised, the image vector is then united with the prefix vector in an exceedingly separate ‘multimodal layer’ that comes once the RNN subnetwork. This separates the priority of modelling the image input, the text input and therefore the combining and interpretation of the encoded inputs.

A screenshot of a cell phone

Description automatically generated

As mentioned, it's common to use a pre-trained model for encryption the image, however equally, this design additionally permits a pre-trained language model to be wont to write in code the caption text input.

Within the merge design, RNNs in result write in code linguistic representations, that themselves represent the input to a later prediction stage that comes when a multimodal layer. it's solely at this late stage that image options area unit want to condition predictions.

There is a unit multiple way that to mix the 2 encoded inputs, like concatenation, multiplication, and addition, though experiments by brandy Tanti, et al. have shown addition to figure higher.

Generally, Marc Tanti, et al. found the merge design to be more practical compared to the inject approach. The success of the merge model for the write in coder-decoder design suggests that the role of the continual neural network is to encode input instead of generate output.

This is a departure from the common understanding wherever it's believed that the contribution of the continual neural network is that of a generative model.

If the RNN had the first role of generating captions, then it might have to be compelled to have access to the image so as to understand what to get. This doesn't appear to be the case as together with the image into the RNN isn't usually helpful to its performance as a caption generator.

Coming back to our original task, the most model is delineating as follows;

We will describe the model in 3 parts:

* Photo Feature Extractor. this can be a 16-layer VGG model pre-trained on the ImageNet dataset. we've pre-processed the photos with the VGG model (without the output layer) and can use the extracted options foretold by this model as input.
* Sequence Processor. this can be a word embedding layer for handling the text input, followed by an extended STM (LSTM) continual neural network layer.
* Decoder (for lack of a more robust name). each the feature extractor and sequence processor output a fixed-length vector. These area unit integrated along and processed by a Dense layer to form a final prediction.

The photograph Feature Extractor model expects input photograph options to be a vector of four,096 parts. These area unit processed by a Dense layer to supply a 256-part illustration of the photograph.

The Sequence Processor model expects input sequences with a pre-defined length (34 words) that area unit fed into Associate in Nursing Embedding layer that uses a mask to ignore cushioned values. this can be followed by Associate in Nursing LSTM layer with 256 memory units.

Both the input models manufacture a 256-part vector. Further, each input models use regularization within the kind of five hundredth dropout. this can be to cut back overfitting the coaching dataset, as this model configuration learns in no time.

The Decoder model merges the vectors from each input models victimization Associate in Nursing addition operation. this can be then fed to a Dense 256 somatic cell layer so to a final output Dense layer that creates a soft max prediction over the complete output vocabulary for ensuing word within the sequence.

The operate named define\_model () defines and returns the model able to be match.

To get a way for the structure of the model, specifically the shapes of the layers (PHOTOES)

We additionally produce a plot to examine the structure of the network that higher helps perceive the 2 streams of input. Next comes model fitting,

Fitting the Model: Now that we all know a way to outline the model, we will match it on the coaching dataset.

The model learns quick and quickly overfits the coaching dataset. For this reason, we are going to monitor the ability of the trained model on the holdout development dataset. once the ability of the model on the event dataset improves at the tip of Associate in Nursing epoch, we are going to save the total model to file.

At the tip of the run, we will then use the saved model with the most effective ability on the coaching dataset as our final model.

We can try this by process a Model Checkpoint in Keras and specifying it to observe the minimum loss on the validation dataset and save the model to a file that has each the coaching and validation loss within the computer file name.

We can then specify the stop within the decision to fit () via the call-back’s argument. we tend to should additionally specify the event dataset in fit () via the validation\_data argument.

We will solely match the model for twenty epochs, however given the quantity of coaching information, every epoch could take half-hour on fashionable hardware. After the outline of the model, we will get a concept of the entire range of coaching and validation (development) input-output pairs. The model then runs, saving the most effective model to .h5 files on the means.

**TRAINING WITH PROGRESSIVE LEARNING:**

If you get a slip-up like 'memory error', This section is for people who don't have enough memory to coach the model as represented within the antecedently. By 'enough memory', we tend to mean either cloud computing resources or having AN excessive quantity of RAM, since deep learning models have a thumping quantity of knowledge, it can't be handled by all laptop systems.

That is why we tend to are proposing another nevertheless adaptive answer by the name of progressive learning. The coaching of the caption model will assume you've got a great deal of RAM. The code within the previous section isn't memory economical and assumes you're running on an outsized EC2 instance with 32GB or 64GB of RAM. If you're running the code on a digital computer of 8GB of RAM, you cannot train the model.

A workaround is to use progressive loading. The first step is we tend to should outline a operate that we will use because the knowledge generator. We will keep things terribly straightforward and have the info generator yield one photo’s price of knowledge per batch. this may be all of the sequences generated for a photograph and its set of descriptions.

The operate below data\_generator () are the info generator and can take the loaded matter descriptions, ikon options, tokenizer and liquid ecstasy length. We will see that we tend to our job the create\_sequence () operate to make a batch price of knowledge for one ikon instead of a whole dataset. this implies that we tend to should update the create\_sequences () operate to delete the “iterate over all descriptions” for-loop.

We currently have just about everything we want. Note, this can be an awfully basic knowledge generator. The large memory saving it offers is to not have the unrolled sequences of train and take a look at knowledge in memory before fitting the model, that these samples (e.g. results from create\_sequences ()) are created pro re natal per ikon. Some offhanded ideas for additional rising this knowledge generator include: Randomize the order of photos every epoch. Work with a listing of ikon ids and cargo text and ikon knowledge pro re nata to chop even additional back on memory.

Yield quite one photo’s price of samples per batch. I have tough with these variations myself within the past. You can saneness check a knowledge generator by job it directly, running this saneness check can show what one batch price of sequences seems like, during this case forty-seven samples to coach on for the primary ikon. Finally, we will use the fit\_generator () operate on the model to coach the model with this knowledge generator.

In our final year project, we'll discard the loading of the event dataset and model checkpointing and easily save the model once every coaching epoch. (you will, you'll, you'll be able to) then return and load/evaluate every saved model once coaching to seek out the one we tend to very cheap loss that you just can then use within the next section. That’s it. you'll currently train the model victimization progressive loading and save a large amount of RAM. this could even be a great deal slower.

A screenshot of a cell phone

Description automatically generatedA screenshot of a cell phone

Description automatically generated

The next step once coaching a model is to gauge it.

With this the implementation section of our project involves AN finish.

Next, we'll be evaluating and testing our model.

# Testing and Evaluation

Evaluate model: Once the model is match, we will assess the ability of its predictions on the holdout take a look at dataset. We will assess a model by generating descriptions for all photos within the take a look at dataset and evaluating those predictions with a regular price operate.

First, we want to be able to generate an outline for a photograph employing a trained model. This involves passing within the begin description token ‘startseq ‘, generating one word, then job the model recursively with generated words as input till the tip of sequence token is reached ‘endseq ‘ or the utmost description length is reached. The operate below named generate\_desc () implements this behaviour and generates a matter description given a trained model, and a given ready ikon as input. It calls the operate word\_for\_id () so as to map AN whole number prediction back to a word. We will generate predictions for all photos within the take a look at dataset and within the train dataset. The operate below named assess\_model () can evaluate a trained model against a given dataset of ikon descriptions and ikon options.

The particular and foreseen descriptions are collected and evaluated together victimization the corpus cheese score that summarizes however shut the generated text is to the expected text. BLEU scores are employed in text translation for evaluating translated text against one or a lot of reference translations. Here, we tend to compare every generated description against all of the reference descriptions for the photograph. we tend to then calculate cheese scores for one, two, three and four accumulative n-grams. BLUE scores: Before moving additional, let’s have a quick detail concerning our analysis live. BLEU, or the Bilingual analysis Understudy, could be a score for examination a candidate translation of text to at least one or a lot of reference translations. Although developed for translation, it is often wont to assess text generated for a set of tongue process tasks. The Bilingual analysis Understudy Score, or cheese for brief, could be a metric for evaluating a generated sentence to a reference sentence. A perfect match ends up in a score of one point zero, whereas an ideal match ends up in a score of zero.0. The score was developed for evaluating the predictions created by automatic computational linguistics systems. it's not excellent, however will provide five compelling benefits:

* It is fast and cheap to calculate.
* It is simple to know.
* It is language freelance.
* It correlates extremely with human analysis.
* It has been wide adopted.

The cheese score was planned by Kishore Papineni, et al. in their 2002 paper “BLEU: a technique for Automatic analysis of Machine Translation “. The plan or technique cultivates by counting matching n-grams within the candidate translation to n-grams within the acknowledged text, wherever one1-gram or unigram would be every token and a written word comparison would be every word combine. The comparison is created notwithstanding ordination. The enumeration of matching n-grams is changed to make sure that it takes the prevalence of the words within the reference text under consideration, not gratifying a candidate translation that generates Associate in Nursing abundance of affordable words. this is often cited within the paper as changed n-gram preciseness. The score is for scrutiny sentences, however a changed version that normalizes n-grams by their prevalence is additionally planned for higher evaluation blocks of multiple sentences.

A perfect score isn't potential in apply as a translation would get to match the reference specifically. this is often not even potential by human translators. the amount and quality of the references accustomed calculate the cheese score implies that scrutiny scores across datasets will be difficult. In addition to translation, we are able to use the cheese score for alternative language generation issues with deep learning ways such as: Language generation, Image caption generation, Text account, Speech recognition and far additional.

The Python linguistic communication Toolkit library, or NLTK, provides Associate in Nursing implementation of the cheese score that you simply will use to judge your generated text against a reference. Sentence cheese Score: NLTK provides the sentence\_bleu () perform for evaluating a candidate sentence against one or additional reference sentences.

The reference sentences should be provided as an inventory of sentences wherever every reference could be a list of tokens. The candidate sentence is provided as an inventory of tokens. Corpus cheese Score NLTK additionally provides a perform known as corpus\_bleu () for conniving the cheese score for multiple sentences like a paragraph or a document. The references should be specified as an inventory of documents wherever every document could be a list of references and every various reference could be a list of tokens, e.g. an inventory of lists of lists of tokens.

The candidate documents should be specified as an inventory wherever every document could be a list of tokens, e.g. an inventory of lists of tokens. Cumulative and Individual cheese Scores The cheese score calculations in NLTK permit you to specify the coefficient of various n-grams within the calculation of the cheese score.

This gives you the flexibleness to calculate differing types of cheese score, like individual and additive n-gram scores. Let’s take a glance.

Individual N-Gram Scores: An Individual N-gram score is that the analysis of simply matching grams of a selected order, like single words (1-gram) or word pairs (2-gram or bigram). The weights are specified as a tuple wherever every index refers to the gram order. To calculate the cheese score just for 1-gram matches, you'll be able to specify a weight of one for 1-gram and zero for two, three and four (1, 0, 0, 0).

Cumulative N-Gram Scores: Cumulative scores talk over with the calculation of individual n-gram scores in any respect orders from one to n and coefficient them by conniving the weighted mean value.

By default, the sentence\_bleu () and corpus\_bleu () scores calculate the additive 4-gram cheese score, additionally known as BLEU-4. The weights for the BLEU-4 are 1/4 (25%) or zero.25 for every of the 1-gram, 2-gram, 3-gram and 4-gram scores. It is common to report the additive BLEU-1 to BLEU-4 scores once describing the talent of a text generation system. Coming back to our project, The NLTK Python library implements the cheese score calculation within the corpus\_bleu () perform. the next score near to one point zero is healthier, a score nearer to zero is worse.

We can place all of this alongside the functions from the previous section for loading the information. we have a tendency to initial ought to load the coaching dataset so as to organize a Tokenizer in order that we are able to cypher generated words as input sequences for the model. it's essential that we have a tendency to cypher the generated words victimization precisely the same coding theme as was used once coaching the model. We then use these functions for loading the check dataset. Running the instance prints the cheese scores. We can see that the scores match at intervals and shut to the highest of the expected vary of a skillful model on the matter. The chosen model configuration is by no suggests that optimized.

Generate New Captions: Almost everything we want to get captions for entirely new images is within the model file. We additionally want the Tokenizer for coding generated words for the model whereas generating a sequence, and therefore the most length of input sequences, used after we outlined the model. We can exhaust code the most sequence length. With the coding of text, we are able to produce the tokenizer and reserve it to a file in order that we are able to load it quickly whenever we want it with no need the whole Flickr8K dataset. an alternate would be to use our own vocabulary file and mapping to integers perform throughout coaching. We can produce the Tokenizer as before and reserve it as a pickle file tokenizer. pkl.

We can currently load the tokenizer whenever we want it while not having to load the whole coaching dataset of annotations. Now, let’s generate an outline for a brand-new photograph. we'll generate an outline for its victimization our model. Download the photograph and reserve it to your native directory with the file name “example.jpg “. First, we tend to should load the Tokenizer from tokenizer.pkl and outline the most length of the sequence to get, required for artefact inputs.

A screenshot of a cell phone

Description automatically generatedA screenshot of a social media post

Description automatically generated

Next, we tend to should load the ikon we tend to that to explain and extract the options. We might do that by re-defining the model and adding the VGG-16 model to that, or we will use the VGG model to predict the options and use them as inputs to our existing model. we'll do the latter and use a changed version of the extract\_features () operate used throughout knowledge preparation, however custom-made to figure on one ikon. We can then generate an outline victimization the generate\_desc () operate outlined once evaluating the model.

# Conclusions

In this project, we have enforced a deep learning approach for the captioning of pictures. The sequent API of Keras was used with TensorFlow as a backend to implement the deep learning design to attain an efficient cheese score of zero point five eight for our model. The Bilingual analysis Understudy Score, or cheese for brief, may be a metric for evaluating a generated sentence to a reference sentence. an ideal match leads to a score of one.0, whereas an ideal pair leads to a score of 0.0. within the future, the authors area unit engaged on alternating Pretrained icon Models to enhance the feature extraction of the model. Also, the authors area unit reaching to improve bring home the bacon higher performance by victimization word vectors on a way larger corpus of knowledge like news articles and alternative on-line sources of knowledge. The configuration of the model was tuned, however alternative alternate configurations will be trained to examine for improvement within the performance of the image captioning model.

## Future Work

Many deep learning-based strategies are projected for generating automatic image captions within the recent years. supervised learning, reinforcement learning, and GAN primarily based strategies are unremarkably employed in generating image captions. each area visual perception (beholding, seeing} and multimodal space will be employed in supervised learning-based strategies. the most distinction between visual perception and multimodal space happens in mapping. Visual space-based strategies perform specific mapping from pictures to descriptions. In distinction, multimodal space-based strategies incorporate implicit vision and language models. supervised learning-based strategies are more classified into Encoder-Decoder architecture-based, integrative architecture-based, Attention-based, linguistics concept-based, Stylized captions, Dense image captioning, and Novel object-based image captioning. Encoder-Decoder architecture-based strategies use an easy CNN and a text generator for generating image captions. Attention-based image captioning strategies concentrate on totally different salient components of the image and come through higher performance than encoder-decoder architecture-based strategies. linguistics concept-based image captioning strategies by selection concentrate on totally different components of the image and may generate semantically wealthy captions. Dense image captioning strategies will generate region-based image captions. conventionalized image captions specific numerous emotions like romance, pride, and shame. GAN and RL primarily based image captioning strategies will generate numerous and multiple captions. MSCOCO, Flickr30k and Flickr8k dataset are common and widespread datasets used for image captioning. MSCOCO dataset is extremely massive dataset and everyone the pictures in these datasets have multiple captions. Visual order dataset is especially used for region-based image captioning. totally different analysis metrics are used for measurement the performances of image captions. cheese metric is nice for little sentence analysis. ROUGE has differing types and that they will be used for evaluating differing types of texts. METEOR will perform AN analysis on numerous segments of a caption. SPICE is best in understanding linguistics details of captions compared to different analysis metrics. though success has been achieved in recent years, there's still an oversized scope for improvement. Generation primarily based strategies will generate novel captions for each image. However, these strategies fail to find distinguished objects and attributes and their relationships to some extent in generating correct and multiple captions.

Additionally, to the present, the accuracy of the generated captions for the most part depends on syntactically correct and numerous captions that successively have confidence powerful and complex language generation model. Existing strategies show their performances on the datasets wherever pictures ar collected from identical domain. Therefore, acting on open domain dataset are a remarkable avenue for analysis during this space. Image-based factual descriptions don't seem to be enough to get high-quality captions. External information will be additional so as to get engaging image captions. supervised learning wants an oversized quantity of labeled information for coaching. Therefore, unattended learning and reinforcement learning are a lot of widespread in future in image captioning.

# References

Brunel University Library (2013) *Harvard Referencing Guide.* Available at: http://www.brunel.ac.uk/\_\_data/assets/pdf\_file/0020/161471/Harvard-Guide.pdf (Accessed: 18 November 2013)

P. Shah, V. Bakrola and S. Pati, "Image captioning using deep neural architectures," 2017 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS), Coimbatore, 2017, pp. 1-4.

L. Chen, H. Zhang, J. Xiao et al., “SCA-CNN: spatial and channel-wise attention in convolutional networks for image captioning,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6298–6306, Las Vegas, NV, USA, June-July 2016.

Neural Networks, 1988. Image reconstruction by neural net processor. 1, p.423.

Zeng, Z. and Li, X., 2019. Application of human computing in image captioning under deep learning. Microsystem Technologies,

Personal Reflection

Reflection on Project

I would have implemented all possible techniques for image captioning and plucked up the courage to do even more difficult tasks than this. I would have tried to put an effort in research side and maybe improvised or enhanced an already existing algorithm to better accuracy.

Personal Reflection

I believe I could have put more effort on my project and implemented all possible techniques for the current problem. I Have learned to be the best version of oneself, if we wish to succeed in our career.

Appendices

More relevant material

The complete code for this project has been submitted separately as a zip file.