

A Project Report
on
IMAGE CAPTION GENERATOR USING MACHINE
LEARNING

*Submitted in the Partial Fulfillment of the Requirements
for the award of*

Bachelor of Technology
in
Electronics & Communication Engineering

By

Divas Gupta(20185116)
Priyanka Soni(20185053)
Dipesh Sharma Poudel(20185105)

Under the guidance of
Dr. Vinay Kumar Srivastava
Professor



Department of Electronics & Communication Engineering
Motilal Nehru National Institute of Technology Allahabad
Allahabad – INDIA

**Department of Electronics & Communication Engineering
Motilal Nehru National Institute of Technology Allahabad
Allahabad – INDIA**

CERTIFICATE

This is to certify that the work contained in the thesis titled “**Image caption generator using machine learning**”, submitted by **Divas Gupta, Priyanka Soni and Dipesh Sharma Poudel** in the partial fulfillment of the requirement for the award of Bachelor of Technology in Electronics and Communication Engineering to the Electronics and Communication Engineering Department, Motilal Nehru National Institute of Technology, Allahabad, is a bonafide work of the students carried out under my supervision.

Date: 25 Nov 2021

Place: Prayagraj

Dr. Vinay Kumar Srivastava
Professor
ECE Department
MNNIT, Allahabad

<Similarity Index Certificate from IPR Cell>

Acknowledgement

We take this opportunity to express our deep sense of gratitude and heart felt thanks to our project supervisor, **Dr. Vinay Kumar Srivastava**, Department of Electronics & Communication Engineering, Motilal Nehru National Institute of Technology, Allahabad for his constant guidance and insightful comments during the course of the work. We shall always cherish our association with him for his constant encouragement and freedom to thought and action rendered to us throughout the work.

We are also thankful to our colleagues and friends for their constant support. Finally, we deem it a great pleasure to thank one and all that helped us directly or indirectly in carrying out this work.

Date:25 Nov 2021

Place:Prayagraj

Divas Gupta(20185116)

Priyanka Soni(20185053)

Dipesh Sharma Poudel(20185105)

Abstract

Generating textual descriptions from the image is major goal of our project. This app combines computer vision concept and natural language. If we can do well in this work, we can use natural language processing technology to understand the world through pictures. We plan to use data sets: Flickr8K, Flickr30K or MSCOCO.

For a deeper understanding of this we, decided to use modern caption generator. An image generator based on our neural network is done in Python through Keras machine learning library.

We have identified four major components of our work:

(R1) data collection; (R2) Convolution Neural Network (CNN) as embedded; (R3) Recurrent Neural Network (RNN) as a decoder; (R4) Sentence Generation and testing. BLEU-4 score selected to evaluate the quality and accuracy of the captions produced.

We have evenly distributed the four categories described above between our team and each member made equal work to further the project. We have successfully completed the implementation of all four components and we are able to train our network in Google Colab.

We also tried to build an end to end project. For this purpose, we designed a flask web app which contains button to upload the image and Predict a button to generate the caption.

Table of Contents

Certificate	i
Similarity index Certificate from IPR Cell	ii
Acknowledgement	iii
Abstract	iv
List of Figures	vii
Abbreviations	viii

Chapter 1: Introduction

1.1 Introduction	1
1.2 Motivation	2

Chapter 2: Literature Survey

2.1 Introduction to ML	3
2.2 Method in ML	3
2.3 Algorithms in ML	6
2.4 Applications of ML	8
2.5 Limitations of ML	10

Chapter 3: Methodology and Architecture

3.1 Data Sources	11
3.2 Data Pre-processing	12
3.3 Convolution Neural Network	13
3.4 Recurrent Neural Network	14
3.5 Loss Function	17

Chapter 4: Result and Performance Evaluation

4.1 Evaluation Metrics	18
4.2 Result	21

Chapter 5: Conclusion

5.1 Conclusion 23

5.2 Future Scope of work 24

References 25

List of Figures

Fig 1.1	Image Caption Generator Pipeline
Fig 2.1	Flow chart of supervised learning
Fig 2.2	Unsupervised learning
Fig 2.3	Pipeline of Reinforcement learning
Fig 2.4	Example of decision Tree
Fig 3.1	Screenshot of initial dataset
Fig 3.2	Dataset after Pre-processing
Fig 3.3	VGG16 architecture
Fig 3.4	Structure of LSTM Cell
Fig 3.5	Captioning model
Fig 3.6	Flow chart of the process
Fig 4.1	Example used for human evaluation in machines
Fig 4.2	BLEU scores for three possible candidate outcomes
Fig 4.3	Three candidate sentences rating using BLEURT
Fig 4.4	BLEURT's data production process with pre-existing metrics
Fig 4.5	Result summary and BLEU Score
Fig 4.6	Some of our accurate captions generated
Fig 4.7	Some of our wrong and partially wrong captions

Abbreviations

AI	Artificial Intelligence
ANN	Artificial Neural Network
BLEU	Bilingual Evaluation Understudy
BLEURT	Bilingual Evaluation Understudy with Representations from Transformers
CNN	Convolution Neural Network
DBM	Deep Boltzmann Machine
EM	Expectation Maximization
LSTM	Long Short Term Memory
RNN	Recurrent Neural Network

Chapter 1

Introduction

1.1 Introduction

Initially, it was considered a very uphill task that a computer could describe an image. With advancement of Deep Learning Techniques, and large volumes of data available, it is possible for us now to make models that can generate textual description i.e., captions describing the image. In this picture caption model based on the semantics, the text associated with photos should be displayed to our liking kind and if possible in native languages also. It has a profound effect on the helping others in the world, for example visually impaired can better understand the content of images in front of them and in turn avoid accidents or to get approximate idea about what is in front of them.

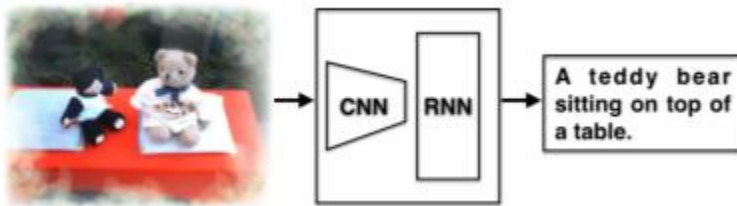


Fig 1.1 Image Caption Generator Pipeline

Our project does this task by using deep neural networks. The framework which we include in our project is combination of a convolution neural network (CNN) which is succeeded by a recurrent neural network (RNN). It gives an English sentence as an output from an image given as input. By observing and learning the knowledge from the image and its corresponding pairs, the method can actually produce image captions that are usually semantically descriptive and grammatically correct to some extent.

1.2 Motivation

In our project, we are doing image-to-sentence construction. This application combines vision and natural language. If we can do well in this task, we can then use natural language processing technology to understand the world in pictures. In relation to that moreover, we present a way of paying attention, which it knows to see what the word means in the picture, and so on summarize the relationship between the objects in the picture. This will be a powerful tool for using non-formatted image data, which controls all data in the earth.

The aim of our project is to develop a web-based environment interface for users to find image description as well to make a separation system to separate images according to their meaning. It can also do the job of SEO(Search Engine Optimization) simple which is as complex as they have to take care of it and check large amounts of data.

This technique can also be used to check any suspicious activities that are happening around by captioning the screenshot of CCTV footage from the cameras in Malls or any other highly secure areas. We can train our model specifically of images of person carrying some sort of weapon. After training model, we can generate caption and connect it to audio device .If any suspicious object will be carried, it will immediately inform the security officials of that area.

These are some of the motivation points which encouraged us to pursue this project and do the needful and add some unique features to it.

Chapter: 2

Literature Survey

2.1 Introduction to ML

Machine Learning (ML) has been developed from Artificial Intelligence, a field of computer science. Machine Learning (ML) is a multi-disciplinary, mathematical compilation and computer science algorithms widely used in predictable analysis and classification. In recent decades, the proliferation of Artificial intelligence (AI) has become a broad and exciting field in computer science as technology prepares machines for human performance, and aims to train computers to solve real-world problems with a high level of success. As we see the growth of science and technological advances AI systems are now able to learn and improve by using prior knowledge without explicit help code when exposed to new data. Ultimately it leads to machine learning technology (ML) that uses learning algorithms to read from available data. Machine Learning uses data mining techniques to extract information from large-sized databases. ML Methods and Data Mining scan data from end to end to detect hidden patterns within the database. Machine learning and data mining algorithms are invested in various fields such as Computer networking, tourism and tourism industry, financial forecasting, telecommunications industry and power forecasting and more.

2.2 Methods used in Machine Learning

Over the years a large number of ML algorithms have been introduced. Only some of them can solve the problem so they replaced the other one. There are mostly used 3 types of ML algorithms for example supervised learning, unsupervised and reinforcement learning.

2.2.1 Supervised Learning

Contains a given set of input variables (training data) pre-labeled and targeted data. It uses a variable input to generate a map function to map the required output to the corresponding input. The parameter adjustment process continues until the system obtains the appropriate level of accuracy regarding the teaching data.

In other words Supervised learning is a machine learning method that maps out output to the desired inputs based on a pair of output inputs. Considering the work from the training data labeled which includes a set of training examples.

Supervised Learning

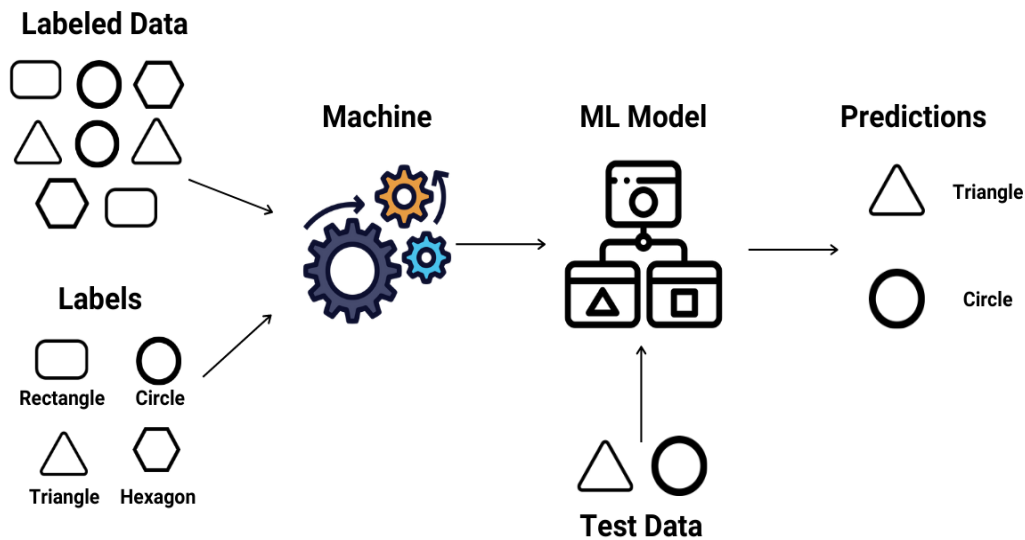


Fig 2.1: Flow chart of supervised learning

The supervised learning algorithm analyzes data given in the form of training and generates activity which we desired in the form of a function, which can be used to map new examples. The right mapping will allow the algorithm to accurately determine class labels in very new situations which are never seen before. This requires a learning algorithm to perform normally from training data to situations that do not appear in a “rational” way.

2.2.2 Unsupervised Learning

In this algorithm we only have training data rather than outcome data. That input data does not have a previous label. It is used for classifiers by identifying existing patterns or collections in input databases. In other words Unsupervised learning is a form of machine learning that looks at patterns that have never been seen in a data set that does not have pre-existing labels and has little human supervision. In contrast to supervised reading that often uses human-written data, Unsupervised Learning, also known as self organization, allows for modeling opportunities for overcrowding. It forms one of the three main stages of machine learning, as well as supervised and reinforced reading.

The two main methods used in Unsupervised learning are the principal component analysis and cluster(group) analysis.

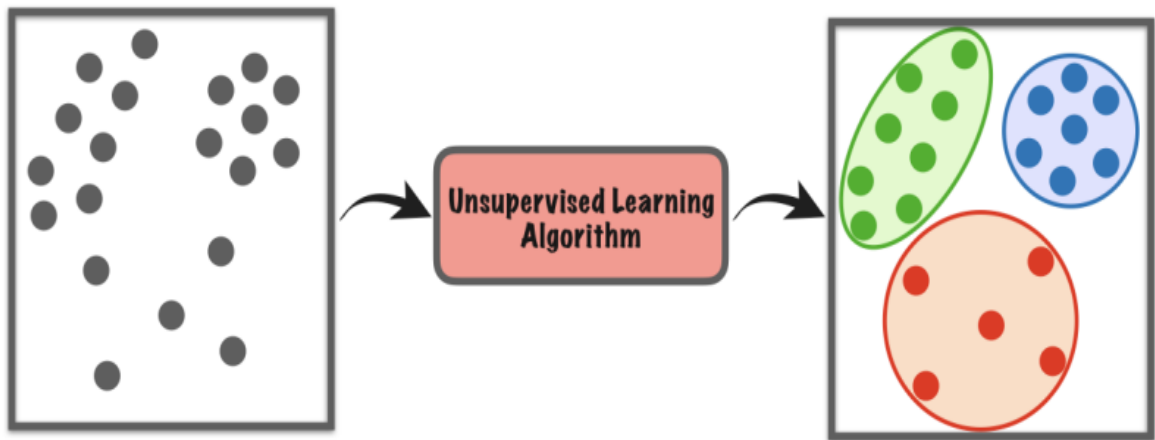


Fig 2.2 : Unsupervised learning

2.2.3 Reinforcement Learning

Using this algorithm machine is trained to map the action to a specific decision which is why the prize or response Symbols are generated. The machine is trained to detect the most rewarding actions by rewards and punishment using previous sensations.

It is a machine learning area about how software agents should take steps in place to develop a vision for the accumulated reward. Reinforcement learning is one of the three basic methods of machine learning, next to supervised reading and non-supervised learning.

Reinforced Learning differs from supervised learning that do not requires paired output/inputs, and require less appropriate steps that need to be clearly adjusted.

Nature is often referred to in the Markov (MDP) decision-making process, because many reinforcement learning algorithms in this context use flexible editing.

Reinforcement Learning

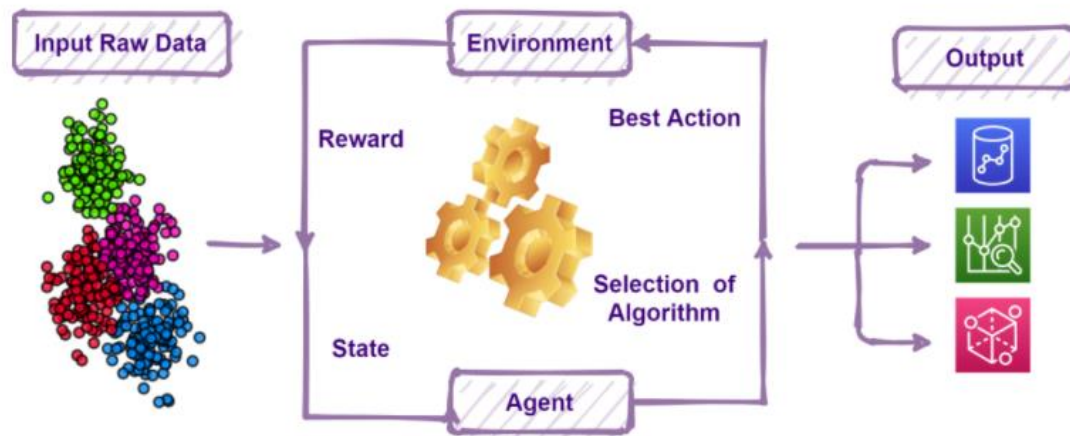


Fig 2.3 Pipeline of Reinforcement Learning

2.3 Algorithm of Machine Learning

2.3.1 Regression Algorithm

In Regression algorithms predictions are modeled by establishing the relationship between variables using error rate. Continuous value is predicted by the Regression strategy. Variables can be value, temperature. Regression algorithms are a type of Supervised algorithms. If we talk about features of this algorithm is that it pays attention on the relationship between the target output and the input features to predict the amount of new data. Algorithms based on regression produce output values on the basis of input features of the data which is given to the system. The algorithm builds a model on training data features and uses that trained model to predict the amount of new data.

Popular areas of regression algorithms are as follows:

- Linear, Quadratic Regression algorithm
- Normal Distribution of Small Squares
- Multivariate Adaptive Regression Splines
- Logistic Regression
- Moderately smooth scatter structure
- Step-by-Step Regression

2.3.2 Decision Trees

Decision trees are most often used in classification problems like logistic regression. They split attributes in two or more groups by sorting them on the basis of their values. Each tree consists of nodes and branches. Attributes of the clusters are represented by

each node and branch represents its value. Pre-pruning and post-pruning are some techniques to improve their accuracy. The most well known algorithms using decision tree are:

- Iterative Dichotomized 3
- Chi squared Automatic Interaction Detection
- C5.0 and C4.5 (different versions of a powerful approach)
- Decision Stump
- Classification and Regression Tree
- Conditional Decision Trees

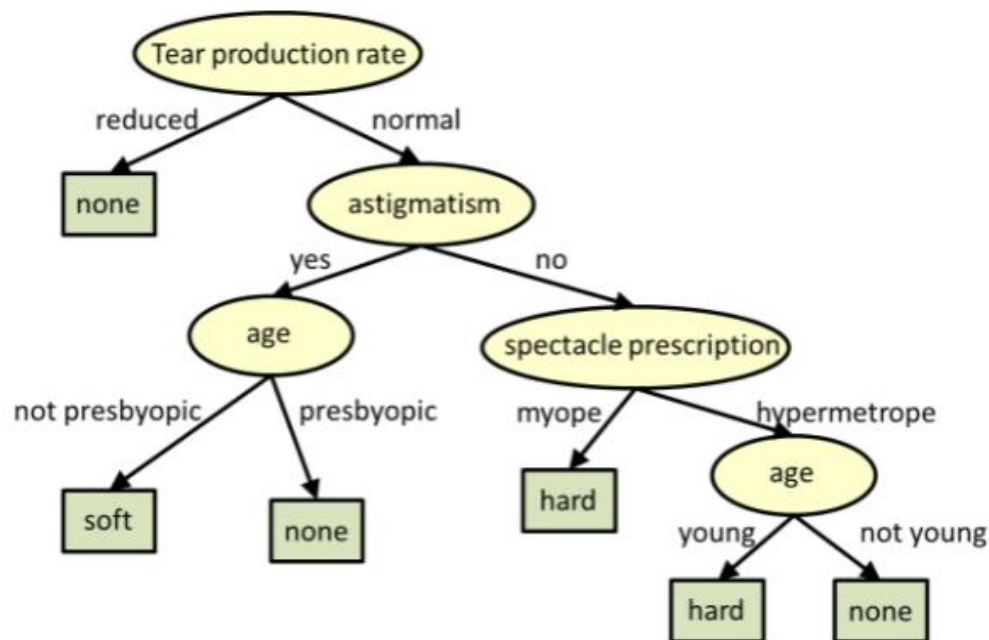


Fig 2.4 Example of Decision Tree

2.3.3 Bayesian Algorithms

Machine Learning is a variety of Computer Science fields like math and algorithm. The statistics control and measure uncertainty and are represented by Bayesian algorithms based on the theory of probability and Bayes theory.

The most common Bayesian algorithms are:

- Bayesian Belief Network (BBN)
- Gaussian Naive Bayes
- Naive Bayes

2.3.4 Clustering Algorithms

This algorithm divides objects into different types of collections. Divides objects into groups where each subset sets share a certain similarity. It is an unsupervised learning method and its methods are categorized as a domain or network clustering and partition clustering. The most popular clustering algorithms are that are used are listed below:

- K Means
- Expectation Maximization (EM)
- K Medians
- Hierarchical-Clustering

2.3.5 Deep Learning Algorithms

Deep Learning techniques are the further advancements made in Artificial Neural Networks. They are more complex neural networks containing lot of hidden layers are large in size. The widely used algorithms for deep learning are:

- Deep Belief Networks
- Deep Boltzmann Machine (DBM)
- Convolution Neural Network (CNN)
- Recurrent Neural Network (RNN)

2.4 Applications of ML

Many industries that work with large amounts of data have recognized the importance of machine learning technology. By obtaining information from this data usually in real time, organizations are able to perform more efficiently or earn more profit than competitors. Other industries that use machine learning are:

Military Applications:

- Inimical energy monitoring
- Monitor friendly forces and equipment
- Observation of military theater or battlefield
- War damage assessment

- Detection of nuclear, biological, and chemical attacks

Nature and environmental Applications:

- Microclimates
- Forest fire detection
- Flood detection
- Agricultural accuracy

Health related Applications:

- Remote monitoring of life data
- Monitor and monitor doctors and patients within the hospital
- Drug administration
- Adult assistance

Home Applications:

- Home automation
- A place with tools
- Automatic meter reading

2.5 ML Shortcomings:

- Each small application requires special training.
- Requires a large amount of training data, organized and structured.
- Learning should be monitored regularly, Training data should be marked.
- Requires long time offline / bulk training.
- Do not learn increasingly or in an interactive manner , or in groups, in real time.
- Low transfer learning capacity, module usability, and integration.
- Systems and training models are opaque, making it very difficult to find bugs.
- Performance cannot be processed or verified at the ‘long tail’.

Chapter 3:

Methodology and Architecture

3.1 Data Sources

We have identified three most used images captions for training data sets in the Computer Vision research domain - COCO data set , Flickr8k and Flickr30k .These databases contain 123,000, 31,000 and 8,000 captions pictures with annotations in sequence and each image has a label with 5 different meanings. Currently, Flickr8k data set containing a small number of images is used as our own the main data source because other two contains more number of images and we have limited storage and calculation capacity. In addition, we also used a Flickr8k clean data separator opened source by Andrej Karpathy as part of our submission Database. This split data modified the original Flickr8k text data in small, discarded non-numeric letters and split data into train, validation, and sub-test sets.

114051287_dd85625a04.jpg#0	A boy dressed in soccer attire and holding his shoes getting out of a car .
114051287_dd85625a04.jpg#1	A boy in a red soccer strip is holding his boots in his hand whilst stepping out of
114051287_dd85625a04.jpg#2	A boy in glasses is wearing a red shirt .
114051287_dd85625a04.jpg#3	A child getting out of the car wearing soccer shoes .
114051287_dd85625a04.jpg#4	A young boy gets out of the van and prepares his shoes for wear during a soccer ga
1141718391_24164bf1b1.jpg#0	A bridge through high green plants , a man and a woman on it .
1141718391_24164bf1b1.jpg#1	A man and a woman are crossing over a rope bridge with greenery all over them .
1141718391_24164bf1b1.jpg#2	A man and a woman are walking across a rope bridge .
1141718391_24164bf1b1.jpg#3	A man and a woman crossing a suspension bridge in a tropical setting .
1141718391_24164bf1b1.jpg#4	Woman and man walking across wooden rope bridge with a caution sign beside it .
1141739219_2c47195e4c.jpg#0	A family gathered at a painted van
1141739219_2c47195e4c.jpg#1	A girl climbing down from the side of a bright blue truck while others watch .
1141739219_2c47195e4c.jpg#2	A man is helping a girl step down from a colorful truck whilst a woman and three c
1141739219_2c47195e4c.jpg#3	A very colorful bus is pulled off to the side of the road as its passengers load .
1141739219_2c47195e4c.jpg#4	Two women and four children standing next to a brightly painted truck .
1142283988_6b227c5231.jpg#0	A blond woman poses with a person in a pink costume .
1142283988_6b227c5231.jpg#1	A smiling woman holds a person dressed in a pig costume .
1142283988_6b227c5231.jpg#2	A young woman hugs a young man who 's wearing a pink costume .
1142283988_6b227c5231.jpg#3	Blond embracing young man in pink costume , at event
1142283988_6b227c5231.jpg#4	Two people are hugging and one is wearing a pink-hooded stretch top .

Fig 3.1 Screenshot of initial dataset

3.2 Data Pre-processing

Input data is made up of images and captions, as well so we need to process both images in the correct order CNN network format and text captions to the correct format RNN network structure. As our caption production pipeline uses state-of-the-art CNN network (VGG-16 which will also be discussed in section 3.3), we need to convert images into correct format.

According to PyTorch documents, the pre-trained vgg16 model expects the input images to be standardized inside width $[0, 1]$ and RGB imagery for 3 channels ($3 \times H \times W$), where H and W are expected to be at least equal to size 224.

Therefore, pre-processing of data involves uploading and resizing image data be ($N \times 3 \times 256 \times 256$) size and normal pixel value to be between distance $[0, 1]$ and mean(expected value) = $[0.485, 0.456, 0.406]$ and standard deviation = $[0.229, 0.224, 0.225]$.

For proper functioning of PyTorch word needs to be translated numbers (indeces) to determine the embedding of words, we develops a glossary of dictionary words for common words in training captions, which include specialty tokens like `<startseq>`, `<endseq>`. Some names only appear less than 5 times in all training captions, and in that case everything is presented as a `<unk>` (unknown).

In addition, captions are coded and dictionary described above and stored in a JSON file that can be downloaded in the RNN model over time.

1000268201_693b08cb0e.jpg	child in pink dress is climbing up set of stairs in
1000268201_693b08cb0e.jpg	girl going into wooden building
1000268201_693b08cb0e.jpg	little girl climbing into wooden playhouse
1000268201_693b08cb0e.jpg	little girl climbing the stairs to her playhouse
1000268201_693b08cb0e.jpg	little girl in pink dress going into wooden cabin
1001773457_577c3a7d70.jpg	black dog and spotted dog are fighting
1001773457_577c3a7d70.jpg	black dog and tricolored dog playing with each other
1001773457_577c3a7d70.jpg	black dog and white dog with brown spots are staring
1001773457_577c3a7d70.jpg	two dogs of different breeds looking at each other
1001773457_577c3a7d70.jpg	two dogs on pavement moving toward each other
1002674143_1b742ab4b8.jpg	little girl covered in paint sits in front of paint
1002674143_1b742ab4b8.jpg	little girl is sitting in front of large painted re
1002674143_1b742ab4b8.jpg	small girl in the grass plays with fingerpaints in
1002674143_1b742ab4b8.jpg	there is girl with pigtails sitting in front of rail
1002674143_1b742ab4b8.jpg	young girl with pigtails painting outside in the gr
1003163366_44323f5815.jpg	man lays on bench while his dog sits by him

Fig 3.2 Data after Pre-processing

3.3 Convolution Neural Network

In project, a convolutional neural network (CNN) maps an RGB image to a visual feature vector. The most widely used layer in CNN are: convolution, pooling and fully-connected layers. Also, ReLU(Rectified Linear unit) $f(x)=\max(0,x)$ is a type of activation function which is used to add non-linearity. The ReLU works much quicker than conventional $f(x) = \tanh(x)$ or $f(x) = (1 + e^{-x})^{-1}$. Overfitting can be avoided by using dropout layer. The dropout adjusts through output of each hidden neuron to zero with a probability (i.e., 0.5). The neurons which are dropped out cannot contribute to the forward pass and in back propagation.

The main function of CNN(as encoder) is to extract features from image and encodes the features into vector space which can be given as input to RNN later. VGG-16 and ResNet are common recommended as image encoders. We have chosen to change the a pre-trained VGG-16 model supplied by the PyTorch library. In this work, CNN is used to extract features instead of a typical classifier. As a result, we removed the fully connected layers and layers of large pools at the end of the network. Underneath this new design, is the image matrix for installation has size $N \times 3 \times 256 \times 256$, and output size is $N \times 14 \times 14 \times 512$. In addition, for support for image sizes of various sizes, we added 2d dynamic background for our CNN architecture.

In our image captions, we have disabled the gradient to reduce computational costs. With good planning, we it may get better performance overall.

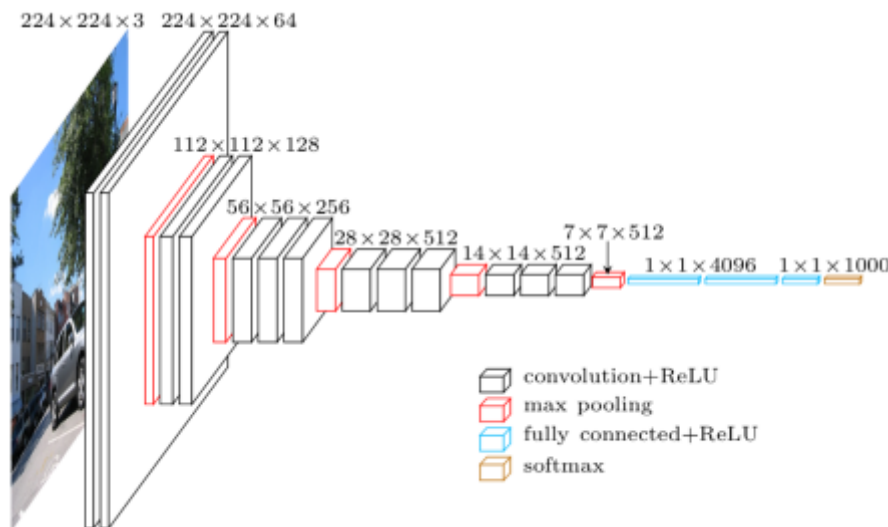


Fig 3.3 VGG16 architecture

3.4 Recurrent Neural Network

LSTM(Long Short Term Memory) which is further advancement in RNN is used in our project as it solves the vanishing gradient problem. Given below the equations for simplified LSTM network at time instant t , given inputs x_t , h_{t-1} , and c_{t-1} :

$$i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + b_i)$$

$$f_t = \sigma(W_{xf}x_t + W_{hf}h_{t-1} + b_f)$$

$$o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + b_o)$$

$$g_t = \phi(W_{xc}x_t + W_{hc}h_{t-1} + b_c)$$

$$c_t = f_t \odot c_{t-1} + i_t \odot g_t$$

$$h_t = o_t \odot \phi(c_t)$$

where $\sigma(x) = (1 + e^{-x})^{-1}$ and $\phi(x) = 2\sigma(2x) - 1$.

LSTM includes hidden unit $h_t \in \mathbb{R}^N$, and also an input gate $i_t \in \mathbb{R}^N$, forget gate $f_t \in \mathbb{R}^N$, output gate $o_t \in \mathbb{R}^N$, input modulation gate $g_t \in \mathbb{R}^N$, and memory cell $c_t \in \mathbb{R}^N$. These additional cells allows the LSTM to train and learn very complex long term temporal dynamics. We can also add depth to the LSTM and stack them on each other.

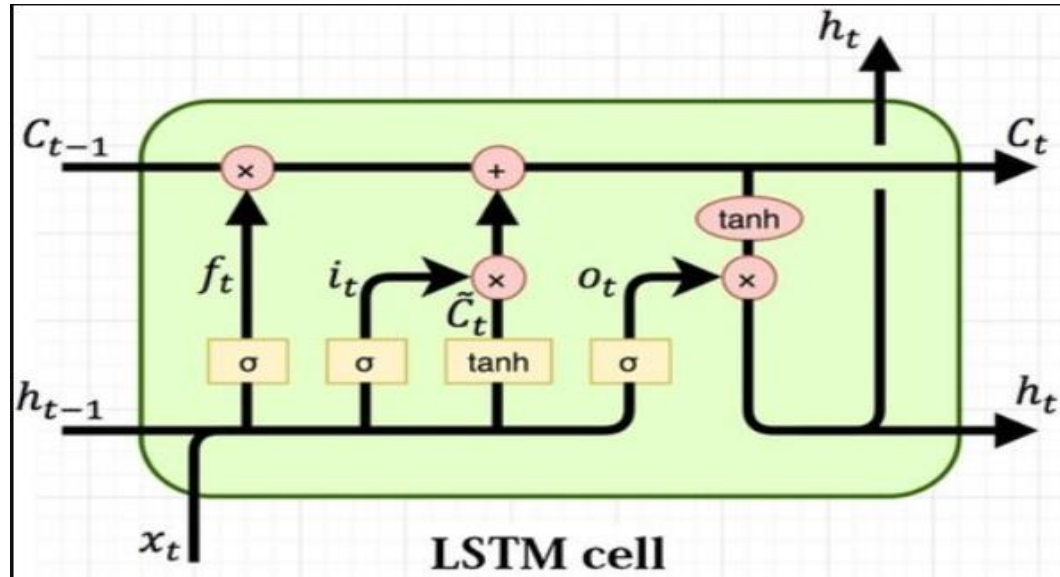


Fig 3.4 Structure of LSTM Cell

The RNN which is used as decoder needs to make captions for the image word by word correctly using Recurrent Neural Network - LSTMs which will be able to produce words in sequence. Input of the decoder is a vector of an encoded image element from CNN as well as coded image captions generated from pre-data processing stage.

The decoder contains a attention module designed and used by ourselves, the LSTM cell module as well four fully integrated layers provided by the PyTorch library for providing initial states of LSTM cell and glossaries.

When we receive encoded images and captions, we pre-filter coded images and captions with a coded key image length in descending order. We only intend to process those image which are encoded and have caption lengths almost greater or equal as compared to number of iterations to increase efficiency and reduction of training time.

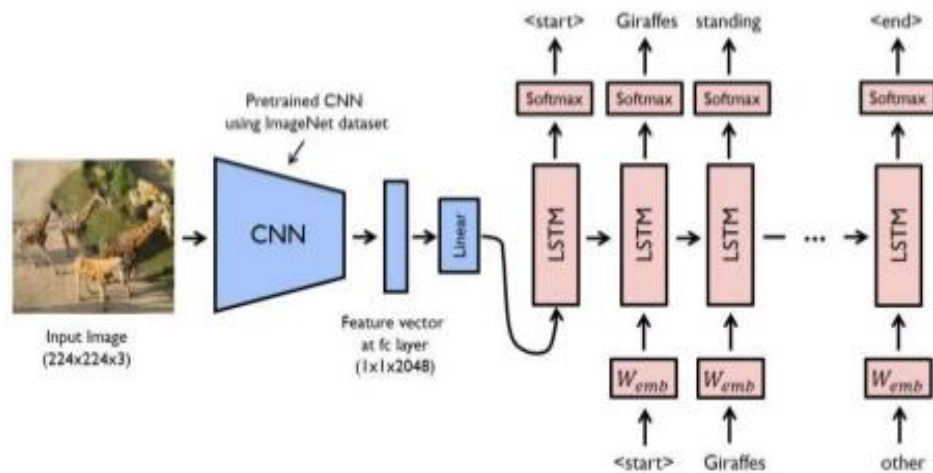


Fig 3.5 Captioning model

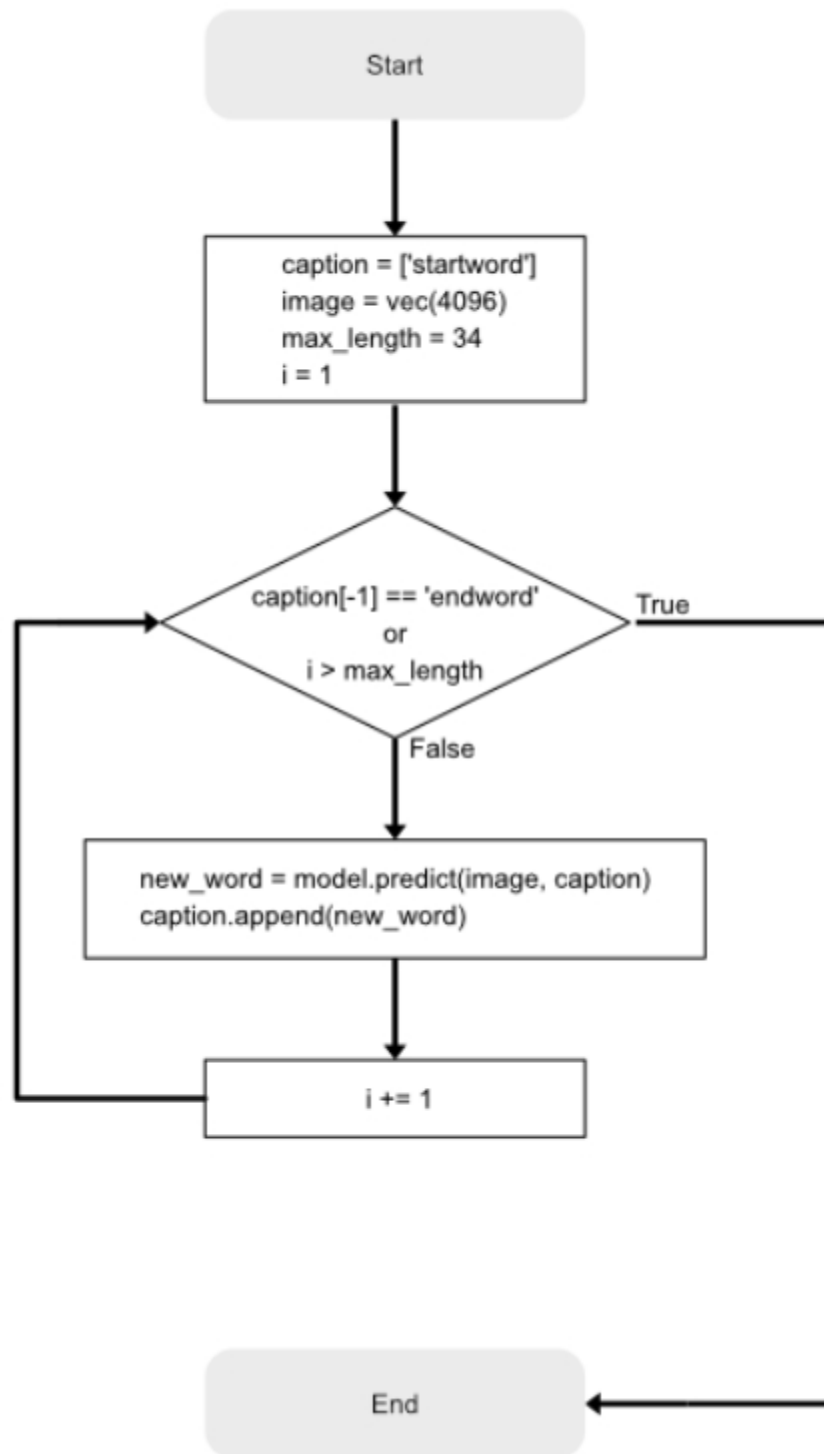


Fig 3.6 Flow chart of the process

3.5 Loss Function

The type of the output of our RNN is a series of probability of word's occurrences in that place, to represent the quality of our RNN output in much understandable numeric form, we decided to use Cross Entropy Loss. This is one of the widely used for accurate measurement of performance for a classification model whole output represents the probability value between 0 and 1.

$$E_{\text{entropy}} = -\sum_n^N t_k^n \ln(p_k^n)$$

where , N= number of classes

$t =$ either 0 or 1

$p_k^n =$ predicted possibility that observation k if of class n

Chapter: 4

Result and Performance Evaluation

4.1 Evaluation Metrics

Text Generation is a subtle domain. Scholars and industry are still struggling to find the right metrics to test the quality of productive models. Every productive work is unique, with its own subtleties and its unique features - chat systems have different target metrics rather than summaries, as does machine translation.

Here are the metrics used in NLP to compare productive or production activities, in which 2 texts should be compared. The metrics I will discuss here can be used for the following activities:

- short or long text production
- Machine Translation
- Summary
- Chat bots and chat systems
- Answers to Questions
- analytical systems
- multimedia systems such as speech2text, photo captions, automatic video copying

In human testing, a piece of the produced text is presented to the commentators, who are tasked with evaluating its quality smoothly and meaningfully. The text is usually displayed side by side with a reference, authorized by a person or from the web.

Input: Bud Powell était un pianiste de légende.
Reference: Bud Powell was a legendary pianist.
Candidate: Bud Powell was a great pianist.

How fluent is the sentence?

○ — ○ — ○ — ○ — ○
not at all *neutral* *very*

Does it accurately convey the meaning of the reference?

○ — ○ — ○ — ○ — ○
not at all *neutral* *very*

Fig 4.1 Example used for human evaluation in machines

The advantage of this method is that it is accurate: people are still friends when it comes to checking the quality of a piece of text. However, this test method can take days easily and involve a large number of people with a few thousand examples, which disrupts the

flow of model development work.

Introducing BLEU

On the contrary, the idea behind automated metrics is to provide a cheap, non-delayed representative of human quality standards. Automatic metrics usually take two sentences as input, candidate and reference, and return points that show how the first one is the same as the last, usually using word spacing. The popular metric is BLEU, which counts the sequence of words in a re-referenced candidate (BLEU result is very similar to accuracy).

Advantages and disadvantages of automated metrics as opposed to those that come with human testing. Automatic metrics are appropriate - can be done electronically in real time throughout the training process (e.g., Tensor board editing). However, they are often inaccurate because they focus on high-level similarities and fail to capture the diversity of human language. Often, there are many effective sentences that convey the same meaning. Quotation-based metrics rely primarily on word combinations that unfairly reward those who are similar in reference to their supernatural, even if they do not take the meaning correctly, and punish other word phrases.

Input: Bud Powell était un pianiste de légende. Reference: Bud Powell was a legendary pianist.	sentence BLEU (0-100)
Candidate 1: Bud Powell was a legendary pianist.	100
Candidate 2: Bud Powell was a historic piano player.	46.7
Candidate 3: Bud Powell was a New Yorker.	54.1

Fig 4.2 BLEU scores for three possible candidate outcomes

Introducing BLEURT

BLEURT is a novel, automated machine-based technique that can capture minor semantic similarities between sentences. Trained in the public collection of ratings and additional ratings provided by the user.

Input: Bud Powell était un pianiste de légende. Reference: Bud Powell was a legendary pianist.	BLEURT
Candidate 1: Bud Powell was a legendary pianist.	1.01
Candidate 2: Bud Powell was a historic piano player.	0.71
Candidate 3: Bud Powell was a New Yorker.	-1.49

Fig 4.3 Three candidate sentences rating using BLEURT



Fig 4.4 BLEURT's data production process with pre-existing metrics

Experiments reveal that pre-training significantly increases BLEURT's accuracy, especially when the test data is out-of-distribution.

We train BLEURT in advance twice, first for the purpose of language matching (as described in the first BERT paper), and then for the collection of NLG test objectives. We then fine-tune the model from the WMT Metrics database, a set of user-provided ratings, or a combination of both.

4.2 Result

We use BLEU to measure the similarity of captions generated by our method and people. BLEU is a popular translation machine matrix that analyzes n-grams integrated events between candidate and reference sentences. Unigram scores (B-1) define translational adequacy, while long n-gram points (B-2, B-3, B-4) represent smoothness.

Summary of our result:

```
Dataset: 6000
Descriptions: train=6000
Vocabulary Size: 7579
Description Length: 34
Dataset: 1000
Descriptions: test=1000
Photos: test=1000
BLEU-1: 0.529691
BLEU-2: 0.277107
BLEU-3: 0.182869
BLEU-4: 0.080318
```

Fig 4.5 Result Summary and BLEU Score



Fig 4.6 Some of our accurate captions generated

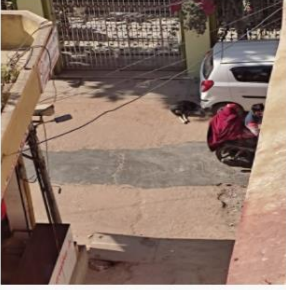

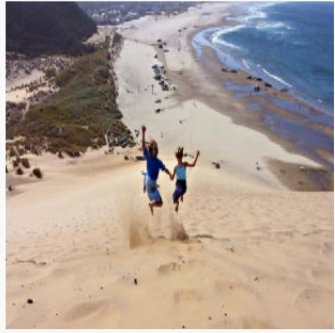

 <p>Result: man in red shirt is sitting on bench</p>	 <p>Result: man in red shirt is riding bike on the dirt</p>
 <p>Result: man is climbing up rock face</p>	 <p>Result: boy in blue shirt is jumping into the water</p>

Fig 4.7 Some of our wrong and partially wrong captions

Chapter: 5

Conclusion

5.1 Conclusion

Automatic photo captions are far from mature again there are many ongoing research projects aimed at more rendering of an accurate image feature and better mathematically sentence production. We ended it with the success we had mentioned in the project proposal, but using a small database (Flickr8k) due to limited computer power.

The first of all in all, we have directly used the pre-trained CNN network as part of our pipeline structure without proper repair, so the network is not as much familiar with this particular training database. So, by trying different CNN networks in advance and empowerment good planning, we expect to achieve high BLEU4 scores. Another potential development is training in a combination of Flickr8k, Flickr30k, and MSCOCO. Usually, a set of different training data on network has apparently, gives the output more accurate. We all agree with this the project introduces our interest in the use of machine learning in Computer Vision and awaits testing more in the future.

Limitations

Our model is designed to capture activities from the image. So, if a completely new activity appears our model will fail and give random results which we have shown in results also. Since, there is no unique caption to a image results will vary according to human requirement and what machine generates. Also, there are grammatical errors in output which makes it sometimes difficult to understand the meaning of sentence.

Also , there are some similar activities like walking ,running, jumping which is difficult for our model to differentiate and also our model is unable to differentiate between the objects in the surrounding. It only considers color of surrounding which sometimes might be deceptive.

5.2 Future Scope of Work

As we mentioned several limitations in our project , project future work can be based on that. Our model is designed to give only one sentence as a caption. We can change it to generate a paragraph to give a variety of different captions. We can combine a wide variety of training data in order to capture as many activities as possible since they are finite in number. Instead of color based surrounding detection , we can train model separately on various surroundings. Because, when there is a white background and a person is walking ,it may be possible that the person is walking on snow, or just there may be a white wall.

We can also develop an android application which will provide user the facility to capture any scenery or image related to different activities with mobile camera and use our model to generate the best suited caption.

References

- [1] Xu, Kelvin, Jimmy Ba, Ryan Kiros, Aaron Courville, Ruslan Salakhutdinov, Richard Zemel, and Yoshua Bengio. “Show, attend and tell: Neural image caption generation with visual attention.” arXiv preprint arXiv:1502.03044(2015).
- [2] Karpathy, Andrej, and Li Fei-Fei. “Deep visual semantic alignments for generating image descriptions” In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3128-3137. 2015.
- [3] Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015: 1-9.
- [4] Vinyals, Oriol, Alexander Toshev, Samy Bengio, and Dumitru Erhan. “Show and tell: A neural image caption generator.” In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3156- 3164. 2015.
- [5] Donahue, Jeffrey, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. “Long-term recurrent convolutional networks for visual recognition and description.” In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2625- 2634. 2015.
- [6] Larochelle H, Hinton G E. Learning to combine foveal glimpses with a third-order Boltzmann machine[C]//Advances in neural information processing systems. 2010: 1243-1251.
- [7] Denil M, Bazzani L, Larochelle H, et al. Learning where to attend with deep architectures for image tracking[J]. Neural computation, 2012, 24(8): 2151-2184.
- [8] Mnih V, Heess N, Graves A. Recurrent models of visual attention[C]//Advances in Neural Information Processing Systems. 2014: 2204-2212.
- [9] Vinyals O, Kaiser , Koo T, et al. Grammar as a foreign language[C]//Advances in Neural Information Processing Systems. 2015: 2755-2763.
- [10] Hermann K M, Kocisky T, Grefenstette E, et al. Teaching machines to read and comprehend[C]//Advances in Neural Information Processing Systems. 2015: 1684- 1692.