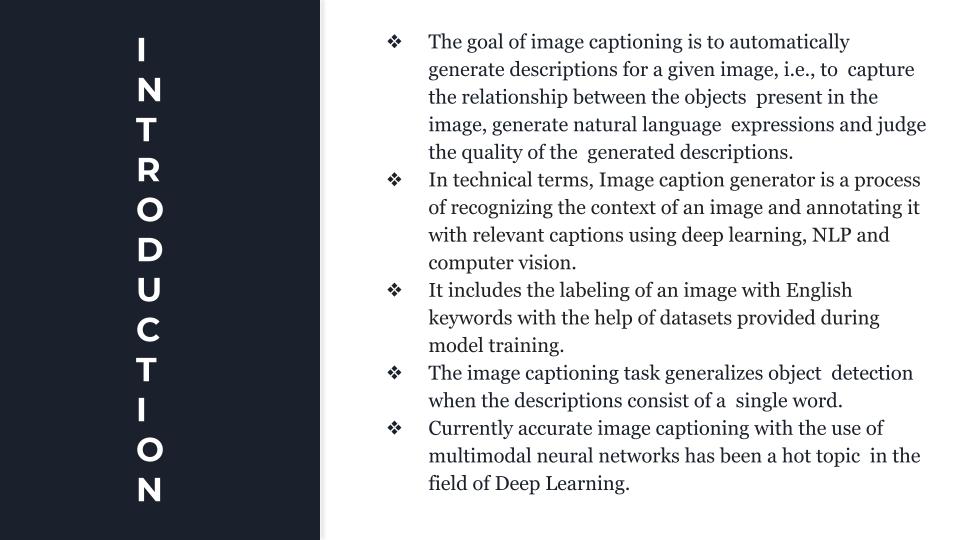
GOOGLE COLAB LINK: <https://colab.research.google.com/drive/1VwFAeswCr6J9us89IQGWntXXPbVZFEmG#scrollTo=3at6TeeLyAEk>



### Abstract

Nowadays, image captioning is a new challenging task that has congregated widespread interest. In this paper, we will discuss and follow some of the core concepts of Image Captioning and its common approaches. Our image captioning task involves generating a concise description of an image in natural language and is currently accomplished by techniques that use a combination of computer vision , natural language processing , and of machine learning methods. In this paper, the model which we have used model generates natural language description of an image. Here, we have used a combination of convolutional neural networks to extract features and then used recurrent neural networks to generate text from these features. If the descriptions just consist of a single word then our image captioning task performs object detection. The obtained results are satisfying and ambitious. We are given with a set of images and prior knowledge about the content and we have to find the correct semantic label for the entire image.

**Keywords**- image captioning, , convolutional neural networks, Long short-term memory, recurrent neural networks

### Introduction

#### Discuss about the problem and possible techniques for proposed system:

The goal of image captioning is to automatically generate descriptions for a given image, i.e., to capture the relationship between the objects present in the image, generate natural language expressions and judge the quality of the generated descriptions. Due to rapid growth and development of technology with the spread of Internet use has leads to the availability of large collections of digital images that are a real fortune for the research community working on machine vision field. But actually the big problem with this huge amount of data is that it is basically unlabelled and this leads to untouched resources. Hence, many recent works are actually dealing with the automatic image caption generation with the intention of giving meaning to these images to be used in advanced artificial intelligence applications. Therefore, it creates a challenging task which will have a high impact in improving research results. And this image captioning field has been around for a decade. But the ability and the productivity of the used techniques was limited at the beginning and they were not strong enough. Therefore, the problem, is apparently more difficult than popular computer vision tasks ,i.e. object detection or

segmentation, where the significance is simply on identifying the different entities present in the image. With recent advancements in training neural networks, the availability of GPU computing power, and large datasets, neural network driven approaches are the most popular choice for handling the caption generation problem.

Anyway, humans are obviously one step ahead at depicting images and building useful and meaningful captions, with or without a particular application context, which renders it an interesting application for IML and explainable artificial intelligence.

Promising technologies include active learning, which was already applied for automating the assessment of image captioning. IML methods to incrementally train, e.g., re-ranking models for selecting the best caption candidate and XAI methods that can improve the user’s understanding of a model and eventually, enable it to provide better feedback for a second IML process. Everything has changed with firstly the free accessibility to huge databases like ImageNet, Flickr 8k

, Flickr 30k and the Microsoft COCO: Common Objects in Context(MS COCO) and secondly the use of encoder decoder framework motivated from the success of neural networks in MV tasks especially the famous architecture deep Convolutional Neural Networks i.e. CNN. This latest skeleton or framework is accepted by many recent researches which has proven her effectiveness as expected.

The CNNs are used as an encoder to extract images features that are next fed to Recurrent Neural Networks (RNNs) for language modelling. The huge disadvantage in this architecture is that it will not examine the structural aspects of the image and automatically generates captions for images by considering the given complete image. To overcome this main limitation, the attention mechanism has been proposed to be incorporated to the encoder-decoder framework. Also it is possible to implement a system that will be able to generate images caption using an encoder/decoder system using CNN and RNN with attention mechanism. In this, Usually, CNN is used as an encoder to extract information from images. Now here, RNN will be used as decoder to transform this present representation into natural language description.

Anyway, the encoder has to squeeze all the input data into a single fixed length vector that is passed to the decoder. Using this vector to compress long and detailed input sequences may lead to loss of information Attention mechanism has been proposed to handle this complexity. In the attention algorithm, an attention vector related to time t is used to replace the fixed length vector obtained from the image encoder CNN. Attention mechanism is assimilated to the encoder-decoder image captioning framework to allow the decoding process to focus on the significant and important details of the

given image at each time step while the output segments are being produced. Many useful improvements are proposed on encoder-decoder structure such as semantic attention and review network. This is achieved through a semantic attention model, which combines semantic concepts and the feature representation of the image/encoding.

#### RELATED WORK:

Caption Recommender System is an integral part of understanding the environment, which has various applications e.g. - subtitle generation, helping visually impaired people to understand their surroundings, storytelling from albums, search using image, etc. Since many years, many different image caption recommendation approaches have been developed. There have been a lot of contributions from the architecture created by the winner of the ILSVRC. Along with the VGG the research made in the field of natural language translation have helped us continuously in bettering the performance in text generation.

Researchers at AI Lab used a Convolution Neural Network for each potential object in the image for producing high-level features of the image. Then a Multiple Instance Learning (MIL) was used for figuring out the best area which matches with each word. This method gave a BLEU score of 22.9% on MS-COCO dataset. The Vinyals came up with a new model

called NIC -Neural Image Caption, Show and Tell model, which was nothing but an encoder RNN which was given input through a CNN model for computer vision. After this a group of researchers took the NIC model and modified it. They used a technique that makes use of images datasets and their corresponding captions to study the inter-modal correlations between natural language and image data. The model used by those researchers was based on a new combination of CNN around image fields, the LSTM or bidirectional RNN over textual descriptions, and a planned aim of putting the two modals together via bimodal embedding. Flickr 30K, Flickr 8K and MSCOCO were the datasets used by them to achieve these bests in business results. Jonathan further modified their model in 2015 when he suggested an idea of a model related to dense captioning in which the model detects each of the different areas of the image and then suggests a group of captions. Chen Wang also suggested a model which makes use of multiple LSTM networks and a deep CNN in the year 2016. Over a period of time there has been enhancements not only in the captioning models but also in score metrics used for evaluating the accuracy of the models. This project has used the BLEU score for evaluation. BLEU - being a standard evaluation metric adopted by many of the groups. Now, new state of the arts metrics has come like CIDER which are

replacing older metrics like BLEU score, etc.

#### APPROACH :

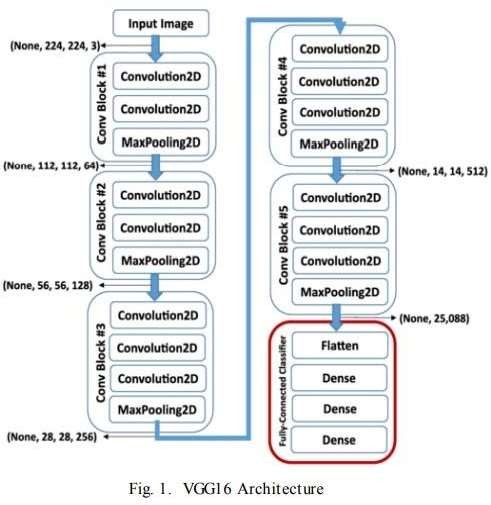
Recent developments in the field of technologies related to image captioning has been the main source of motivation for our research work. The model proposed in this paper has an eventual aim as to predict natural language descriptions for various areas of the image. The research work focuses mainly on obtaining the results for several image captioning models by making use of BLEU score metric and hence comparing the performance of different image captioning models.

Various CNN models such as VGG16, Inception-V3 etc are used for encoding the images and extracting features from the images. Further these encoded images are used with two types of decoders, namely unidirectional LSTM and bidirectional LSTM to obtain the results. We have used greedy search and beam search algorithm to generate the caption from encoded features. The generated caption is then compared with the original caption from dataset on the basis of Bilingual Evaluation Understudy score.

#### Convolution Models Encoders:

This section discusses various convolution models used for the research work. There are two encoders namely, VGG16 and Inception-V3. Each convolution model has been described in brief in the following subsections.

VGG16: VGG16 consists of a 16-layer network for the completion of the task of encoding the image. Out of 16 layers present in the VGG16 network, 3 are dense layers and rest 16 are convolution layers. The architecture of VGG16 is shown below.For the feature extraction to be done on the image, the dimension of the image has to be a 224\*224 image. We have fixed the length of the stride to be 1 for the CNN layer which have filters of size 3\*3. The next step is Max pooling, it is executed using a window size of 2\*2-pixel with a length of stride taken to be 2.



**Inception-V3:** InceptionV3 consists of a 48-layer deep convolutional network for performing the task of encoding the image. InceptionV3 stacks together 11 inception modules each of which consists of convolution and max- pooling layers. For the feature extraction to be done on the image,

the dimension of the image has to be a 229\*229 image. Three fully connected layers of size 512, 1024 and 3 are added to the final concatenation layer.

1. **Decoders** :This section discusses various decoder models used for the generation of captions for images. There are two decoders used in this research work namely, unidirectional LSTM and Bidirectional LSTM. Each type of LSTM network has been described in brief in the following subsections.
2. **LSTM (Long Short-Term Memory):** LSTM have been widely used by the researchers in the areas of text translation, audio to text conversion etc. As in the traditional RNNs, the straight structures are also present in LSTM, but there is a difference in the building manner of the reiterating modules. The main method by which LSTM preserves the past info is by line running on the top of LSTM network which is called as cell states. All of the modules in the network consist of a cell state. These cell states are fed information with the help of different gates. These gates are composed up of sigmoid function -whose value varies between 0 and 1- so it can be decided how much information is to be passed to the next layer. If the value of the sigmoid function is 1, it means the whole of the information is passed to the next cell else if it is 0 then no information is passed. Hence the cell states help the network to maintain the info in the system.
3. **Bidirectional LSTM:** Bidirectional LSTM are an addendum to the conventional LSTMs and can help in significantly enhancing the performance of the model problems related to sequence classification. A Bidirectional LSTM, or bi-LSTM, is a model for sequence processing that consists of 2 LSTMs: one taking the input in a forward direction, and the other in a backwards direction. Bidirectional LSTMs work upon 2 LSTMs in place of one on the sequence provided as input.The first LSTM trains itself on the input sequence as-is and the second LSTM works upon the reversed copy of the input sequence. By using the bi-LSTMs the amount of information available to the network is increased effectively, which helps in enhancing the context available to the algorithm and thus result in complete and faster learning of the model.

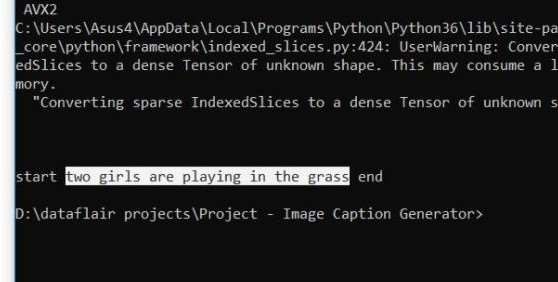
**Dateset Collection** :We have used the Flickr 8kdataset for training and validation purposes. This dataset has been provided by University of Illinois at Arbana-Champaign. The dataset contains 8000 images and for each image it has corresponding 5 descriptions. By default, the dataset is split into two folders, image folder and text folder. For each image the caption is stored along with the respective id as we have a distinctive image-id for every image in the set. The images in the dataset are divided into three parts: Training set, validation set and Test set. Test and validation set consist of 1000 images each whereas the

training set consists of 6000 images. Apart from this there are other datasets also available like MS-COCO and Flickr30k for captioning images but both these datasets have at least 30,000 images and training the model on these datasets requires a lot of power and is computationally very expensive.

**Fig. A image from dataset:**



**Captions generated:**



**Data Preprocessing** :Flickr 8k dataset consists of nearly 6000 train images and for each image we have corresponding 5 descriptions. These text descriptions require some minimal pre-processing before we can use it to train the model. We first loaded the file containing all the descriptions along with their corresponding image id. We looped through the file and created a dictionary which maps each photo identifier to a list containing

textual descriptions for the image. After this we did some cleaning of the textual data in order to reduce vocabulary size. Cleaning of textual descriptions involve: removing punctuations, converting text to lowercase, removing stop words like ‘a’, ‘an’ etc. and removing tokens containing digits. Next step is to create a vocabulary of all the unique words present across all the image descriptions. Finally, for each description which corresponds to an image in training dataset we need to add a ‘’ token at the start of each caption and an token at the end of each caption. The token signifies the start of a sequence while token signifies end of a sequence.

**Feature Extraction:** In our research work, image acts as an input to the decoder network. For training the decoder, the image data must be provided in the form of fixed size vectors. Therefore, each image is converted into a fixed size vector which will then be fed as input to RNN. We use a transfer learning method for extracting features from the images.

For this purpose, we used pre-trained models and its weight trained on larger similar data. We computed the image features using these pre-trained models and saved them in a file. Later we loaded these features and fed them into the neural network as the interpretation of the image given in the dataset.

**Model Training and Evaluation**: For training purposes, we used the Google colaboratory notebook. We trained the decoder model on a batch size of 32 and 64 using Adam optimizer and categorical\_crossentropy as loss function. We used training and validation loss as the metric to evaluate the model after each epoch. We monitored the validation loss of the model during training. When the validation loss of the model improves at the end of an epoch, we saved the model into a file. At the end of the training period, we used the model with best skill on the training dataset as our final model.

#### Need/Motivation:

Image captioning has turned out to be a challenging and important research area including advances in statistical language modelling as well as image recognition. Also, there are various practical benefits on generating captions from images starting from aiding the visually impaired, to enabling the automatic and cost saving labelling of the millions of images uploaded to the internet every now and then. The image captioning field also induces together state of the art models in Natural Language Processing and Computer Vision which are the two of the major fields in Artificial Intelligence. Due to rapid increase in technology and advancement of image classification and object detection, it becomes possible to automatically generates one or more sentences to

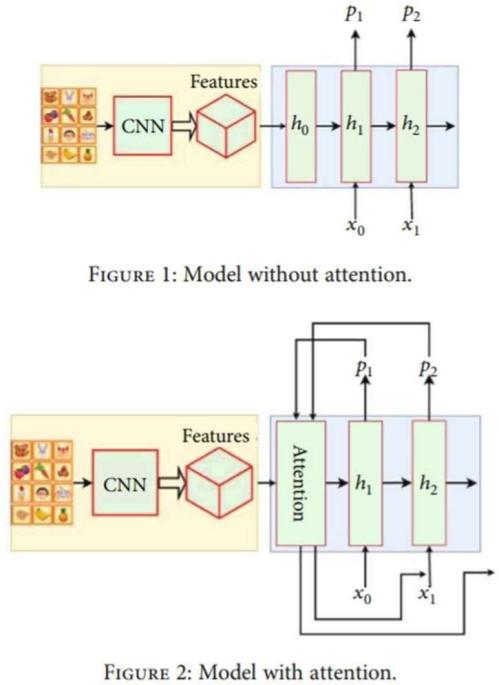
identify the content of an image and making it easy to understand. Other advantages on generating complete and natural image descriptions has large potential effects, such as titles attached to news images, text-based image retrieval, information accessed for blind users descriptions associated with medical images as well as including interaction between human and robot. We know that the remarkable effect produced due to applications in image captioning has a lot important theoretical and practical research value. Hence, image captioning is a more convoluted but a significant and meaningful task in the age of artificial intelligence.

#### Objectives-

* To detect objects on the scene and determine the relationships between them.
* Express the image content correctly with properly formed sentences.
* To automatically describe an image with one or more natural language sentences.

1. **Literature Survey:**
   1. **Automatic Image Captioning Based on ResNet50 and LSTM with Soft Attention.**

Here in this paper, they investigated one single-joint mode, AICRL, for automatic image caption generation using ResNet50 (a CNN) and LSTM (long short-term memory) with soft attention mechanism. AICRL mainly consists of an encoder and a decoder. The model was designed with one encoder-decoder architecture. They adopted ResNet50, a convolutional neural network, as the encoder to encode an image into a compact representation as the graphical features. A language model LSTM was selected as the decoder to generate the description sentence. Meanwhile, they integrated the soft attention model with LSTM such that the learning can be focused on a particular part of the image to improve the performance.

The entire model is fully trainable by using the stochastic gradient descent that makes the training process easy. The investigational results indicate that the proposed model is able to generate good captions for images automatically.

The final purpose of AICRL is to generate the proper description for the given images. For that, the AICRL model is designed with an encoder-decoder architecture based on CNN and RNN. In particular, to extract visual features, they used the ResNet50 network as the encoder to generate a one- dimensional vector representation of the input images. After that, to generate the description sentences, they adopted the LSTM as the language model for the decoder to decode the vector into a sentence. Meanwhile, they have utilized the soft attention in the decoder to authorize the model to selectively focus the attention over a certain part of an image to anticipate the next sentence. They have evaluated the model using several popular metrics such as BLEU, METEOR, and CIDEr. BLEU is an algorithm that measures the precision of an -gram between the generated and reference captions. METEOR (Metric for Evaluation of Translation with Explicit Ordering) is an evaluation metric which was initially used in machine translation. CIDEr measures the similarity of generated captions to their ground truth sentences for evaluating image captioning. This measurement considers the grammaticality and correctness.

They performed an extensive set of experiments to evaluate the effectiveness of the proposed model. They have adopted two different datasets in our experiments including the MS COCO 2014 dataset and Flickr\_8K dataset, which contain the images with their descriptions. The MS- COCO 2014 dataset contains 102739 pictures with their descriptions, five captions for each image, and 20548 testing examples.

* 1. **Hiearchical Attention-Based Fusion for Image Caption with Multi-Grained Rewards**

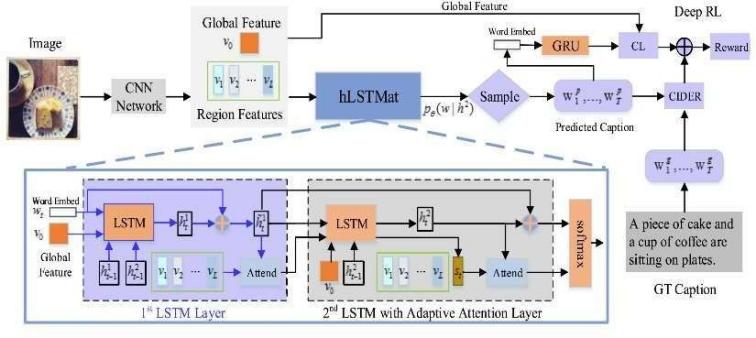
Here in this paper, a Hiearchical Attention fusion model is presented as a paradigm for image caption based on RL, where multi-level feature maps of Resnet are integrated with hierarchical attention. Revaluation network is exploited for revaluating CIDEr score by assigning different weights for each word according to the importance of each word in a generating caption.

* Revaluation network(REN) offers facilitating revaluation reward calculation, which allocates different importance to the generated words in a sentence automatically throughout the RL training phase.
* Scoring network(SN) performs to provide a score as sentence-level reward. It evaluates a generated caption from both the correspondance to the matched ground fact and the discriminates to the unmatched ground fact.

Multi-level feature maps of Resnet are integrated with hierarchical attention, HAF acts as a paradigm for RL-based image caption approach, Moreover, multi-grained rewards are presented in RL phase to revise the proposed HAF. Revaluation network is exploited for reward revaluation by estimating the different importance of each word in a generating caption. The revaluated reward is achieved by weighting the CIDEr scores, where the different masses are calculated from Revaluation network. The revaluated reward can be treated as word-level reward. To obtain benefits from additional unmatched ground truth, Scoring Network(SN) is implemented to provide a score for the generating sentence from a batch of ground truth as sentence -level reward.

The main contributions of this paper are:

* A Hiearchical Attention Fusion (HAF) acts as a paradigm of RL training for image caption. It adopts multiple attention to attend hiearchical visual features of convolution neural networks which take advanatges of multi-level visual matter.



The dataset used here is MSCOCO 2014 caption dataset. For validation of model and offline testing, we use the ‘‘Karpathy’’ splits that have been used extensively for reporting results in most previous works. The split contains 113,287 training images with five captions each, and 5000 images for validation and testing respectively. We filter the vocabulary and drop any word that has counted less than 5, resulting in a vocabulary of 9, 680 words.

* 1. **Camera2Caption: A Real-Time Image Caption Generator.**

In this paper, they followed some of the

main concepts of Image captioning and

|  |
| --- |
| its approaches. Here encoder and |
| decoder based implementation with |
| significant changes and optimizations |
| which enables to run these models on |
| low-end hardware of hand-held devices is |
| used and also compared results |
| corrected using various metrics with |
| state-of-the-art models and analyse why |
| and where our model trained on |

MSCOCO dataset which lacks due to the quality.

In this implementation they followed the

trade-off between computation speed and

similar approach used in Show and Tell

|  |
| --- |
| by introducing an encoder-decoder |
| architecture system. The encoder being |
| pre-trained InceptionV4 Convolution |
| Neural Network by google and the |
| decoder, a Deep Recurrent Neural |
| Network with Long Short Term Memory |
| Cells. Encoder InceptionV4 is used to |
| transform raw images into a fixed length |
| embedding which constitute the |
| convolved features for the images. The |
| decoder here has two phases they are, |
| training and inference. The decoder is |
| mainly responsible for learning the word |
| sequences given in the convolved |

features and original caption.

Initially, the model is trained with flickr30k

dataset with 31,783 images with five

|  |
| --- |
| captions each, but owing to less number |
| of training samples and every training |
| caption beginning with “A man … ” |
| template, but this model failed to |
| generalize, so they shifted to MSCOCO |
| (2014) training dataset with 82780 |
| images, each image with five ground truth |
| captions. For offline purpose they used |

karpathy split3 dataset.

1536.This model supports only JPEG and

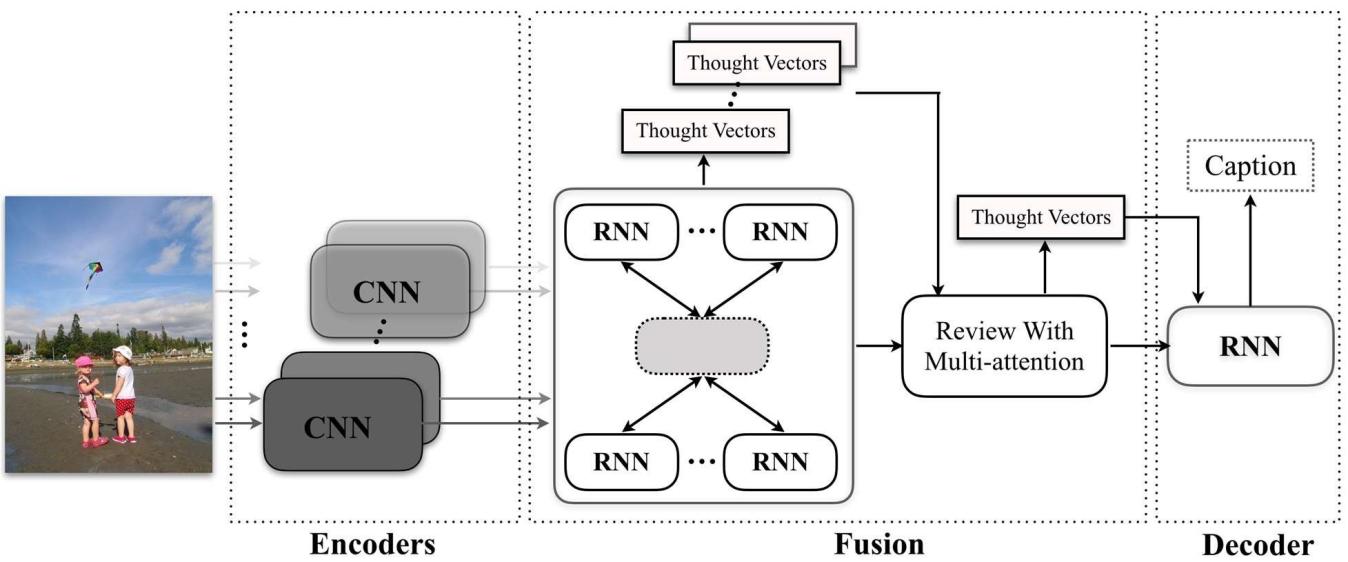
inceptionV4 to encode the image

|  |
| --- |
| they used pre-trained encoder instead of |
| training. They focussed mainly on |
| implementation for quicker caption |
| generation instead of achieving state-of- |
| the-art performance and due to this there |
| exists few obstacles for a justified |
| comparison. Here visual attention is not |
| used, they train and perform inference |
| using a single model, whereas some of |
| other executions use an ensemble to |
| boost their score and differences of |
| dataset splits, due to the lack of a |

During, training they used the average

pooling layer from a pre-trained resulting in a vector of dimension PNG image formats. To avoid over-fitting

standardized split for offline evaluation.



* 1. **Unpaired Image Captioning via Scene Graph Alignments**

In this paper, they have presented a

scene graph based approach for

|  |  |  |  |
| --- | --- | --- | --- |
| unpaired image captioning. The frame- | | | |
| work comprises an image scene graph | | | |
| generator, a | sentence | scene | graph |
| generator ,a scene graph encoder, and a | | | |

sentence decoder. Particularly, we train sentence decoder on the text modality.

the scene graph encoder and the

|  |
| --- |
| To adjust the scene graphs between |
| images and sentences, they proposed an |
| unsupervised feature alignment method |
| that maps the scene graph features from |

the image to the sentence modality.

used to build the data correspondence

between the two modalities. Particularly,

|  |  |  |
| --- | --- | --- |
| given the | unrelated image and | the |
| sentence | scene graphs, they | first |
| encoded them with the scene graph | | |

encoder trained on the sentence corpus.

Then, performed unsupervised cross-

modal mapping for feature level

|  |
| --- |
| alignments with CycleGAN. By mapping |
| the features, the encoded image scene |
| graph is pushed close to the sentence |
| modality, which is then used effectively as |
| input to the sentence decoder to generate |

meaningful sentences.

They proposed a scene graph-based

Here Visual Genome dataset is used to

method that exploits the rich semantic

|  |
| --- |
| information captured by scene graphs. |
| Framework comprises an image scene |
| graph generator, a sentence scene graph |
| generator, a scene graph encoder, a |
| sentence decoder, and a feature |
| alignment module that maps the features |
| from image to sentence modality. They |
| first extracted the sentence scene graphs |
| from the sentence corpus and train the |
| scene graph encoder and the sentence |

decoder on the text modality. images and sentences, CycleGAN is

To adjust the scene graphs between

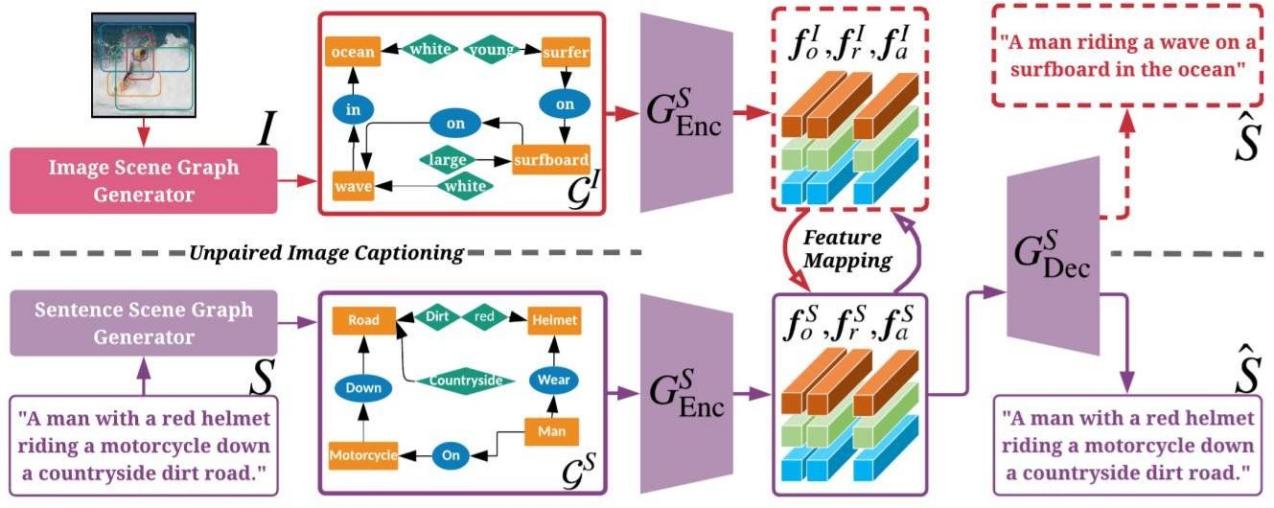
train our image scene graph generator.

|  |
| --- |
| They filter the object, attribute, and |
| relation annotations by keeping those that |
| appear more than 2000 times in the |
| training set. The resulting dataset |
| contains 305 objects ,103 attributes and |
| 64 relations (a total of 472 items). They |
| collected the image descriptions from the |
| training split of MSCOCO and used them |
| as sentence corpus to train the scene |
| graph encoder and the sentence |
| decoder. In pre-processing, they tokenize |

the sentences and convert all the tokens

to lowercase. The tokens that appear less

than five times are treated as (UNK) tokens.



* 1. **TOPIC-GUIDED ATTENTION FOR IMAGE CAPTIONING**

In this paper, they proposed a novel

attention mechanism, called topic-guided

|  |
| --- |
| attention, which integrates image topics |
| in the attention model as a guiding |
| information to help select the most |
| important image features. Moreover, |
| extracted image features and image |
| topics with separate networks, which can |

be fine-tuned jointly in an end-to-end manner during training.

Here, they proposed a topic-guided

attention mechanism that uses the image

|  |
| --- |
| topic as a high-level guiding information. |
| This model starts with the abstraction of |
| image topic, based on image’s visual |
| appearance. Then, the topic vector is fed |

into the attention model jointly with the feedback from LSTM to generate

attention on image visual features and

attributes. The experimental results generate captions that are more

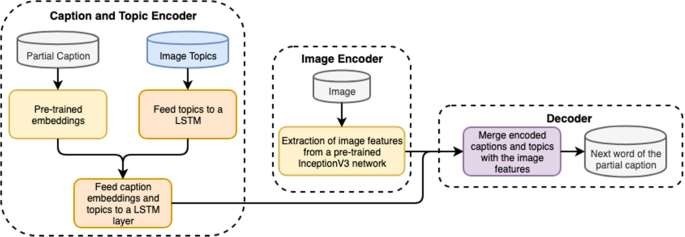
accordant with image’s high-level

demonstrate that this method is able to

semantic content.

The main contributions of this work are:

* They proposed a new attention mechanism which uses image topic as auxiliary guidance for attention generation. The image topic acts like a regulator,



consistent with the general image content.

maintaining the attention

* They proposed a new approach to integrate the selected visual features and attributes into caption generator. This algorithm is able to bring off state-of-the-art performance on the Microsoft COCO dataset.

Here Microsoft COCO dataset is used. For fair comparisons they followed the commonly used split in the previous works: 82,783 images are used for training, 5,000 images for validation, and 5000 images for testing purpose. Some images have more than 5 captions which are discarded for consistency.

For the encoding part:

1. The image visual features v are taken from the last 512 dimensional convolutional layer of the VGGNet.
2. Topic extractor uses the pre-trained VGGNet linked with one fully connected layer which has 80 unites. Its output is the probability that the image belongs to each topic.
   1. **Reference Based LSTM for Image Captioning**

In this paper, training images are

considered as the references and

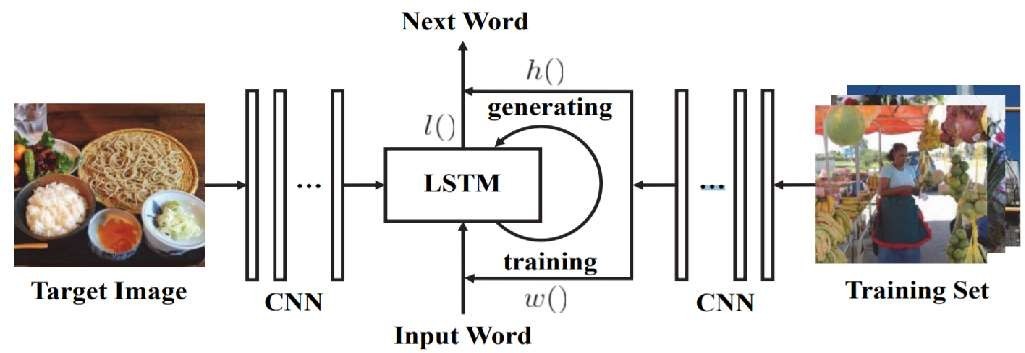
|  |
| --- |
| propose a Reference based Long Short |
| Term Memory(R-LSTM) model, aiming to |
| solve the two problems they are in |
| training phase it is difficult to find which |
| parts of the captions are more essential |
| to the images; and in the caption |
| generation phase, the objects or the |
| scenes are sometimes are unrecognized |
| or misrecognized. In this model when |
| training the model different weights are |
| being assigned to distinct words which |
| enables the network to better learn the |
| key information of the captions. while |
| generating caption, the agreement score |
| is utilized to exploit the reference |
| information of neighbour images, which |
| fixes the misrecognition and make the |
| descriptions more natural. The model |
| proposed here i.e R-LSTM surpass the |
| state-of-the-art approaches on the top |

dataset MS COCO.

This work follows the encoder-decoder

model. But distinct from (Vinyals et al.

2015), the words in a caption are weighted in the training phase according to their relevance to the corresponding image, which well balances the model with the importance of a word to the caption. In the generation phase, they



had taken advantage of the consensus

score (Devlin et al. 2015a) to improve the quality of the sentences. Different from

Mao et al. (2015) who simply used the

consensus score to re-rank the final

|  |
| --- |
| candidate descriptions, consensus score |
| is used in the whole generation process, |
| which means that the decoder takes the |
| neighbours information of the input image |
| into the account. Firstly, the deep VGG- |
| 16 model is employed as the encoder to |
| extract CNN features of the target image |
| and the training images. The weights |
| fixed to each word in the training captions |
| is also calculated. Secondly, the LSTM |
| model is trained using the weighted |
| words and CNN features of the training |
| images, and is adopted as decoder, |
| which takes the CNN features of the |
| target image as input and generates the |
| description words one by one. Here |
| likelihood and the consensus score jointly |
| considered as the evaluation function in |

beam search.

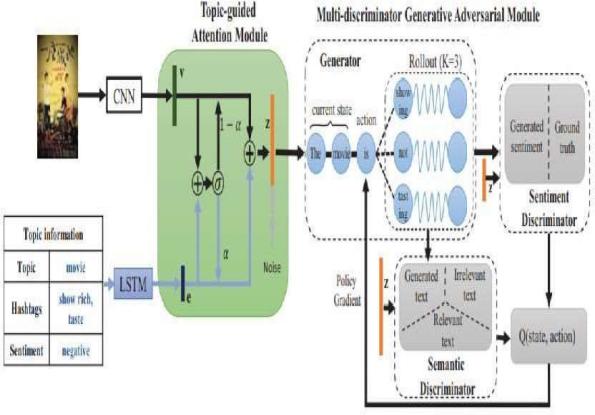
superior over the state-of-the-art

Experimental results on MS COCO

confirmed that the proposed R-LSTM is approaches for image captioning.

# Colloquial Image Captioning

Here,in this paper they consider colloquial image captioning a new task, which, however, is confined in the visual-textual semantic gap, as well as the semantic and the sentiment diversities. To this end, we propose a novel topic-guided multi-discriminator captioning model, termed TGMD-Cap. To facilitate quantitative evaluation, we further construct and release a colloquial image captioning dataset, CIC-dataset, crawled from real-world social network, serving as the first of its kind. Quantitative results show that the proposed TGMD-Cap model outperforms the state-of-the-art approaches under various standard evaluation metrics.

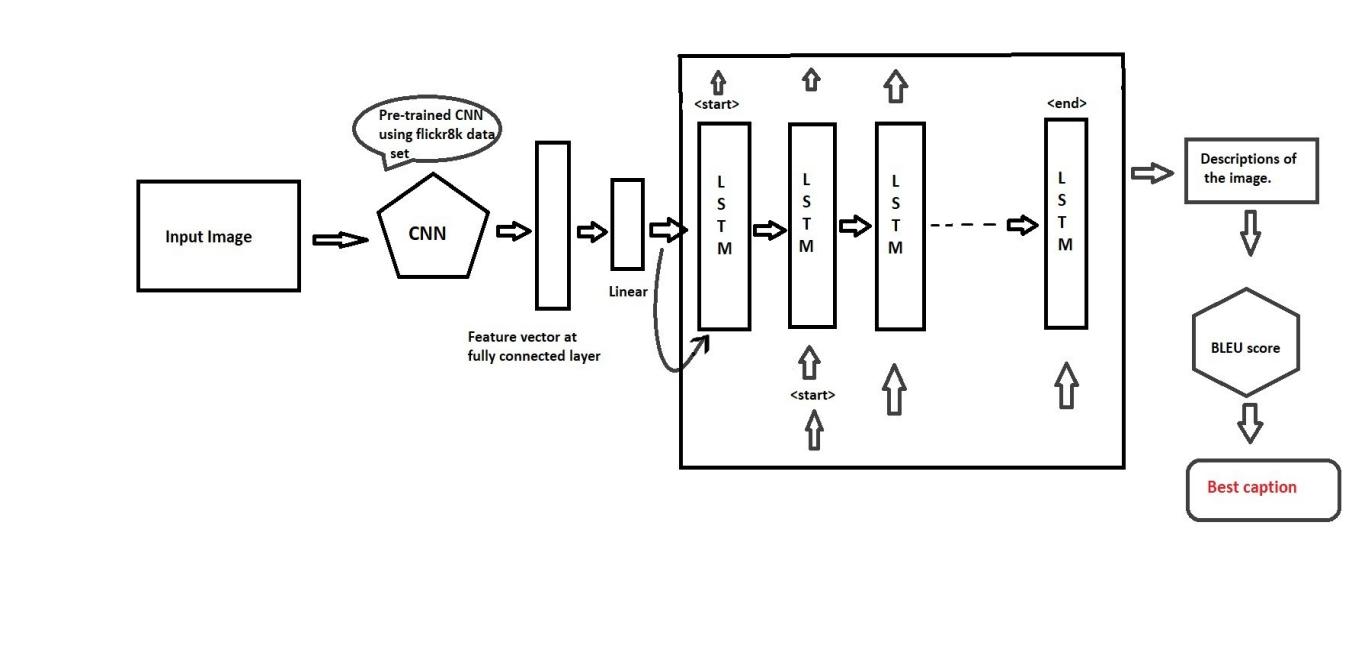
The problems of semantic gap and semantic-and sentiment diversities appear when employing the conventional image captioning methods in colloquial image captioning.

Specially, when on one hand, semantic gap appears between the non-opinion related result of conventional method and the colloquial caption in human view. On the other hand, the expected method should generate the different expressions with diverse semantic contents and sentiments as human-beings.

Driven by the above insights, we proposed a novel topic-guided multi-discriminator captioning model, termed TGMD-Cap, with the topic guided input and the multi-discriminator output for colloquial image captioning, as illustrated . Our main innovations lie in two aspects. Firstly, we construct a topic-guided attention module to compose the effective visual semantic features to reduce the visual-textual semantic gap. Secondly, we design a multi-discriminator generative adversarial module to keep the diversities of the colloquial captions, where two synchronous discriminators, i.e., the semantic discriminator and the sentiment discriminator, are designed for the correct semantic contents and the expected sentiment polarity respectively.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Sno.** | **publication** | **Paper Title** | **Technique** | **Advantages** | **Limitations** |
|  |  | **Hiearchical** |  |  |  |
| **1.** | **IEEE,2020** | **Attention-Based Fusion for Image Caption with**  **Multi-Grained** | **MSCOCO 2014**  **dataset + HAF REN+SN** | **Boost the caption performance significantly.** | **High Intra-class**  **diversity** |
|  |  | **Rewards** |  |  |  |
| **2.** | **IEEE,2017** | **Camera2Caption: A Real-Time Image Caption Generator.** | **Flickr30K dataset**  **+InceptionV4**  **CNN + RNN** | **It transforms raw images into fixed length embedding.** | **This model supports only JPEG and PNG format images.** |
| **3.** | **IEEE,2019** | **Unpaired Image Captioning via Scene Graph** | **Encoder-decoder**  **Framework+ VG**  **+ MSCOCO** | **The model achieves**  **better performance**  **and is more effective. The** | **The task of manually**  **labelling images can be**  **tedious for larger** |
|  |  | **Alignments** | **datasets** | **captions generated are more relevant.** | **datasets.** |
| **4.** | **AAAI,2017** | **Reference Based LSTM for Image Captioning** | **R-LSTM+VGG-16**  **+CNN** | **Surpasses the existing techniques with better accuracy.** | **Assigning weights makes it a little complex** |
|  |  |  |  | **It provides high** |  |
|  |  |  |  | **level semantic** |  |
| **5.** | **IEEE,2018** | **Topic-Guided Attention for** | **Microsoft COCO**  **dataset+ NIC +**  **deep neural** | **content of the**  **image and is instructive for** | **The effectiveness of the**  **caption is not that**  **much good.** |
|  |  | **Image-captioning** | **network** | **guiding the** |  |
|  |  |  |  | **attention to exploit** |  |
|  |  |  |  | **image’s fine** |  |
|  |  |  |  | **grained local**  **information.** |  |
| **6.** | **Hindwai,2020** | **Automatic Image Captioning** | **ResNet50** | **It is able to generate good** |  |
|  |  | **Based on ResNet50 and LSTM with Soft Attention.** | **+LSTM+BLEU+**  **CIDEr + METEOR** | **captions for the**  **images automatically and training process is easier.** | **It’s efficiency is less.** |

**Architecture Diagram:**



Here, we are gonna input an image and that is sent to CNN which is already pre-trained Xception model with flickr8k dataset then we take feature vector of the output from the CNN. We are then sending that to another linear layer where we map it to the input of LSTM. The size of the output from the linear layer would be the embedded size and then we would output the start token (<start>) and that start token will be the input for the next step then start predicting the actual contents of in the image.

e

There is of a difference how we do this at training time versus testing.so during testing for the images that we don’t actually

have correct captions for that what we are going to do is we will feed the the previous output as the next input. During training we are not gonna have that the input to next is the output from the previous. The descriptions of the images are obtained then using the BLEU score the suitable caption is generated.

Xception model is an extension of the inception Architecture which replaces the standard Inception modules with depth wise seperable convolutions. The model is separated in 3 flows

* + Entry flow
  + Middle flow with 8 repetitions of the same block.
  + Exit flow.

# OVERCOMING CHALLENGES IN AUTOMATED IMAGE CAPTIONING

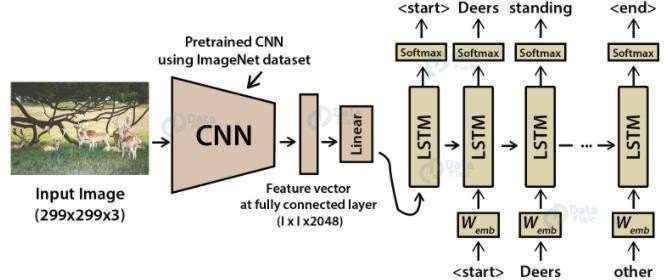
Automatic image captioning remains challenging despite the recent impressive progress in neural image captioning. In the paper appearing at the 2019 Conference in Computer Vision and Pattern Recognition (CVPR), we – together with several other address three main challenges in bridging the semantic gap between visual scenes and language in order to produce diverse, creative and human-like captions. Traditional captioning systems suffer from lack of compositionality and naturalness as they often generate captions in a sequential manner, i.e., next generated word depends on both the previous word and the image feature. To address the issue of lack of naturalness, we introduce another innovation by using generative adversarial networks (GANs) in training the captioner, where a co-attention discriminator scores the “naturalness” of a sentence and its fidelity to the image via a co- attention model that matches fragments of the visual scenes and the language generated and vice versa. The Co-attention discriminator judges the quality of

a caption by scoring the likelihood of generated words given the image features and vice versa. Note that this scoring is local (word and pixel level) and not at a global representation level. The IBM group Research has initiated focused efforts called Code Risk Analyzer to bring security and compliance analytics to DevSecOps. Code Risk Analyzer is a new feature of IBM Cloud Continuous Delivery and a cloud service that helps provision toolchains, automate builds and tests, and control quality with analytics.In the Progress of automatic image captioning and scene understanding will make computer vision systems more reliable for use as personal assistants for visually impaired people and in improving their day-to-day life. The semantic gap in between bridging language and vision points to the need for incorporating common sense and reasoning into scene understanding.

#### PROJECT WORK (a)IDEA FORMULATED

Basically the idea formulated is that CNN will be used for extracting features from the image. We will be using a pre- trained model calsled Xception .LSTM will use the information from CNN and generate a description of the image. Along with this Adam optimization will be used so that the learning process speeds up.

#### ARCITECTURE DIAGRAM

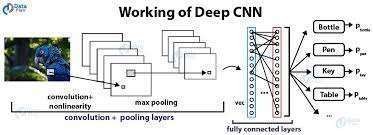


1. **DATASET USED**

There is a wide range of datasets for automatic image description research. The pictures along with textual descriptions in these datasets differ from each other in certain aspects such as in size, the format of the descriptions.

The dataset that we have used is Flickr8K dataset. Flickr8k dataset consists of around 8,000 images that are each paired with five different captions which provide clear descriptions of the salient entities and events.

#### TECHNIQUE CNN:



A convolutional neural network is a type of artificial neural network used in image recognition and processing that is precisely designed to process pixel data

.Basically, CNN is used for image classifications and identifying if the image is of a man, car, animal etc. It scans images and gives important features from the image and combines the feature to classify images.

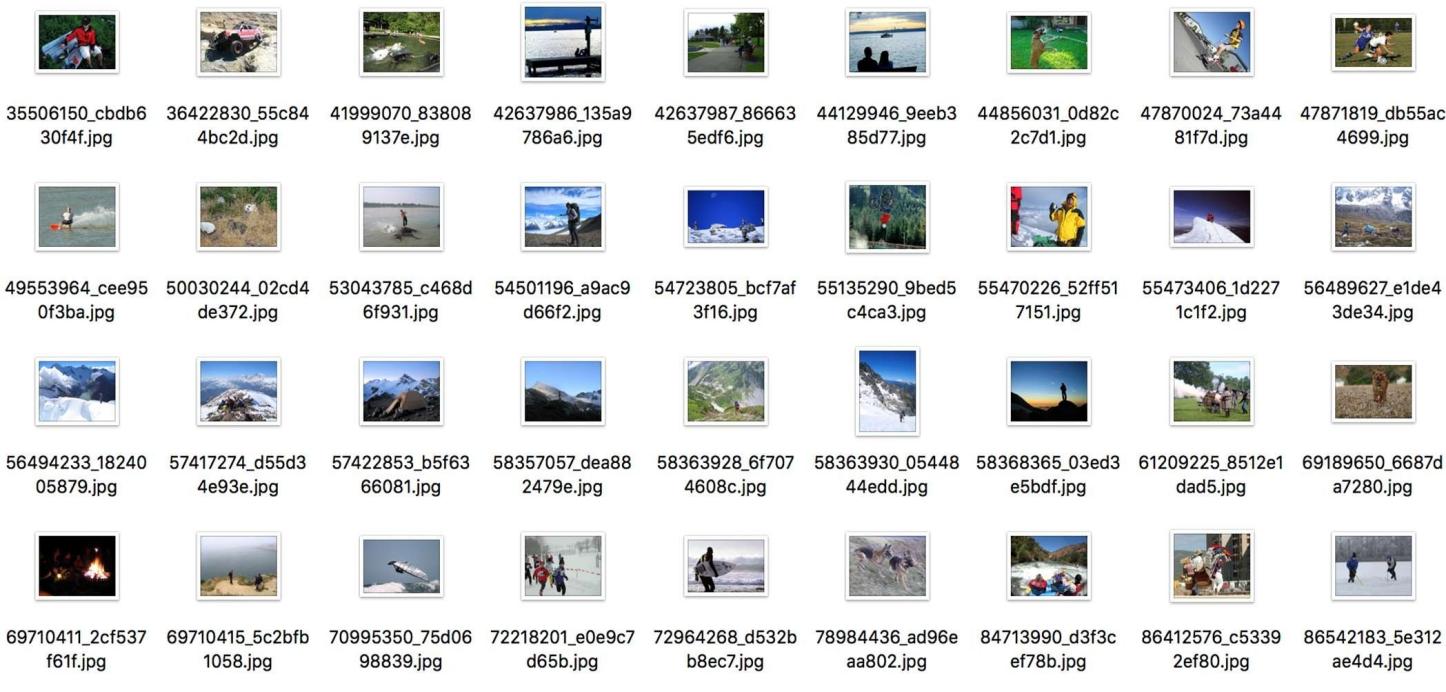
Due to some variations in neural networks that have helped advanced deep learning, and one of those is convolutional neural networks (CNNs). CNNs are most often involved with analysing visual imagery.

Deep learning has made a significant advancement in image processing When the networks have a convolutional layer they are referred to as CNNs. An important property of a CNN is that it can detect image features such as bright, dark, or specific colour spots, edges in various orientations, patterns, and more.

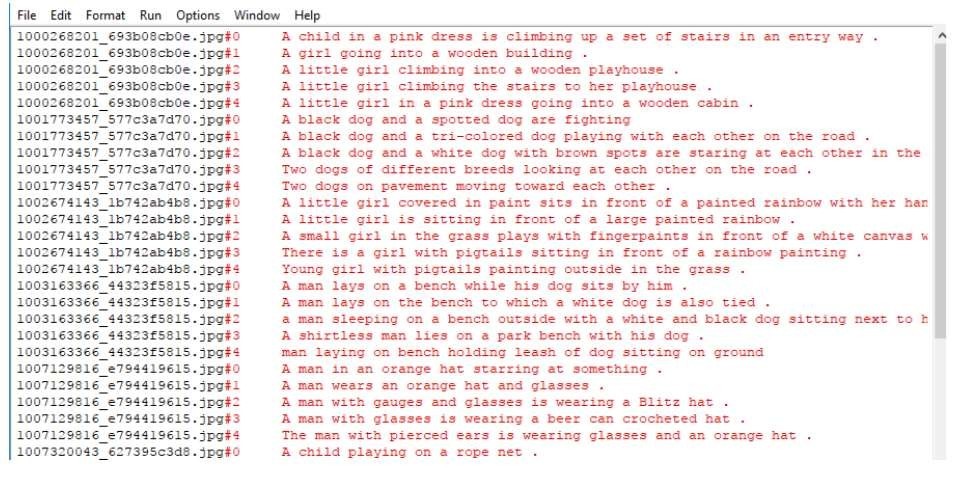
By using more basic features, a CNN can detect more advanced ones such as a human’s ears, a dog’s tail, a person’s eye, or certain shapes. For a regular neural network, this task of detecting advanced features based on the pixels of the input image is very difficult. The features can show up in different positions, different orientations, and they can be different sizes within the image. For example, even if the object looks exactly the same to our eyes .any movement of the

object or camera angle can cause the pixel values to dramatically change

Flickr8k dataset



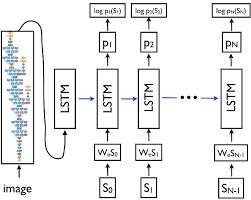
captions



#### Xception:

Xception is a deep convolutional neural network architecture which involves Depth wise Separable Convolutions. Therefore this will be used for generating captions to the image which will also increase the efficiency of the whole model.

#### LSTM



Long short-term memory(LSTM) is an artificial recurrent neural network architecture used in the field of deep learning.

There has been a problem with sequence prediction since a long time now. They are said to be one of the most difficult problems to solve in the data science industry. These include problems such as from predicting sales to finding patterns in stock markets’ data,from language translations to predicting your next word on your iPhone’s keyboard.

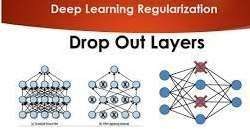
With the recent advancements in data science, it is found that for almost all of these sequence prediction problems, Long short Term Memory networks, also known as LSTMs have been observed as the most effective result.

LSTMs have an edge over conventional feed-forward neural networks and RNN in a number of ways. This is due to

their property of selectively remembering patterns for long durations of time.

It not only processes single data points, but also entire sequences of databecause of its feedback connection.LSTM carries out relevant information throughout the processing of inputs.With a forget gate, it removes non-relevant information.

#### ADAM OPTIMISATION



We have adopted the method of Adam optimization so that the learning process speeds up. Adam optimisation gradually decreases the learning rate, converging more quickly. We use Adam optimization with regularization methods known as dropout. The main advantage of this method is that it prevents all neurons in a layer from synchronously optimizing their weights. Therefore by applying the dropout technique in convolutional layers with a value of 0.5 and 0.3 in the LSTM layers helps to avoid overfitting that quickly happens with a small training set like the Flickr8K dataset

#### BLEU ALGORITHM

For problems, such as summaries, language translations, or captions we use Metrics called BLEU. Full BLEU Bilingual Evaluation Understudy form. It is a test matrix for a given sentence in a reference sentence.

BLEU uses a modified version to compare candidate translation with multiple reference translations. Metric converts easy accuracy as machine translation systems are known for producing more words than reference text. BLEU has often been reported as a good combination of human judgment and remains the standard test for any new test metrics. Apart from this a lot of criticism has been noted. It has been noted that, although legally able to test translations of any language, the BLEU, in its current form, cannot deal with languages without word boundaries. It has been argued that although BLEU has significant advantages, there is no guarantee that the increase in BLEU points is an indicator of improved translation quality. There is a coherent, structured problem with any metaphor based on one or a few translations: in real life, sentences can be translated in many different ways, sometimes without overlap. Therefore, the method of comparing how computer translation differs from a few human translations is flawed. HyTER is another automated MT metaphor that compares multiple translations into the grammatical index defined by human translators going back and forth so that the effort of a person involved in correctly defining the many related ways of interpreting functional translation means HyTER is also a measure.

BLEU output is always a number between 0 and 1. This number indicates the similarity of the candidate text with the reference text, with values very close to 1 representing the same text. The perfect match is 1 and the odds are 0.

Very few people's translations will get 1 point because this will show that the person being baptized is like one of the most trusted versions. For this reason, it

is not mandatory to get 1 point. Because there are so many matching options, adding additional reference translations will increase BLEU points.

. The following are the sections of the BLEU algorithm: For each i to N, calculate the Si values of the gram co- arithmetic for both candidates and reference (Ccand, refs) and the calculation of i -gram from the election (Ccand). Si = Ccand, refs / Ccand… (1) The average price of Si. This is accomplished for the purpose of geometric weights. Weight wi is always maintained for all i (wi = 1 / N for all i). SN = e ^ (SIGN (wi \* log (Si)))… (2) Count the shortest sentence. If the length of the candidate (c) is greater than the length of the candidate (r), then there is no penalty (b = 1). Alternatively, the penalty is logarithmically based on two lengths: b = e ^ (1- (r / c)) if tr… (3) Finally, calculate total points (BLEU points) as the purpose of all scores is multiplied by the shorter penalty . BLEUScore = SN \* b… (4) where there is no penalty (b = 1). Alternatively, the penalty is logarithmically based from a two-dimensional length:

Finally, calculate the total score (BLEU Points) as a description of all scores multiplied by the shorthand penalty. BLEUScore = SN \* b… (4)

To detect changes in BLEU points due to the difference in sentence length, tests were performed with various calculations of T patterns against H. Table 1 shows the test results. The length of T will be longer than H because the addition of another word (defined as X). In the table we also saw that T2 has two patterns that produce the same points. The T3 pattern has an "X" above the T2 pattern, so that the BLEU level drops. In addition, to see some different patterns, then we doa list of possibilities patterns for all T. The results can be

seen in Table 2. If there is a sentence “A B C D” then five types of patterns will be

generated. While the number of adding X is symbolized by Xn.

Table 1 BLEU score comparison based on sentence

patterns.

|  |  |  |
| --- | --- | --- |
| Hypothesis/ Text | Sentence Patterns | BLEU  Score |
| H | {ABCD} | 1 |
| T1 | {ABCD} | 1 |
| T2 | {A B C D X} or {X A B C D} | 0,77 |
| T3 | {A B C D X X} or {X X A B C D} | 0,60 |
| T4 | {A B C X D} or {A X B C D} | 0,59 |
| T5 | {A B C X X D} or {A X X B C D} | 0,46 |
|  |  |  |

Table 2. Possibility patterns from adding “X” words.

|  |  |
| --- | --- |
| Pattern Types | Sentence Patterns |
| Pattern-1 | {A B C D Xn} or {Xn A B C D X} |
| Pattern-2 | {A B C Xn D} or {A Xn B C D} |
| Pattern-3 | {A B Xn C D} |
| Pattern-4 | {A B Xn C Xn D} or {A Xn B Xn C D} |
| Pattern-5 | {A Xn B Xn C Xn D} |

Table 3 contains the results of BLEU scores from the patterns in the Table

2. The value of text it n in the table is the number of adding “X”. If n equal to 0, then there is no addition of “X”

Figure 1 has comparison graph of BLEU score for all patterns. In the

graph can be seen that the as X increases, the value of BLEU score decreases. In addition, the decline is quite drastic. For eg: in Pattern-4 and Pattern-5, adding of “X” can reduce almost half of score, that is from 1 to 0.4

Table 3. BLEU Score for all patterns in Table 2 with adding n-word of ‘X’.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| No | Pattern-1 | Pattern-  2 | Pattern-3 | Pattern-4 | Pattern-  5 |
| 0 | 1 | 1 | 1 | 1 | 1 |
| 1 | 0,7  7 | 0,59 | 0,70 | 0,46 | 0,47 |
| 2 | 0,6  0 | 0,46 | 0,54 | 0,28 | 0,22 |
| 3 | 0,4  7 | 0,35 | 0,42 | 0,17 | 0,10 |
| 4 | 0,3  6 | 0,28 | 0,33 | 0,10 | 0,05 |

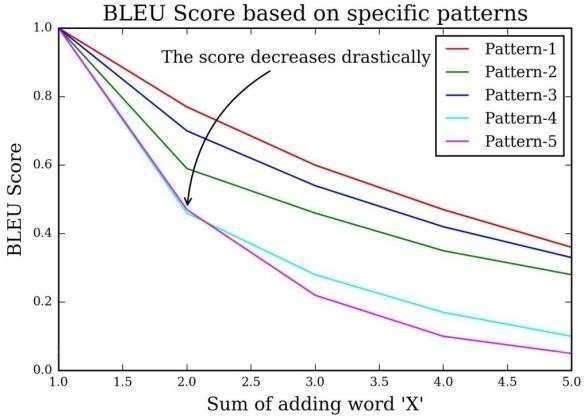
Quite a number of T were same patterned on patterns in Table no.2 These discoveries became the motivation for reducing the length of T with the assumption that, “Thesmaller of difference in length of T and H, the higher of BLEU score”. The sentence reduction in this study uses word removal techniques. The basic idea is to remove several words in T so that the length of T becomes shorter or close to H. The sentence reduction algorithm can be seen in Algorithm-1.

Figure 1. BLEU Score comparison between allpatterns

Table 4. The experimental results using dataset DS-200-R.

|  |  |
| --- | --- |
| Dataset | Accuracy (%) |
| Original BLEU | 77,5 |
| Modified BLEU (MBRTE) | 83,82 |

Techniques

Image Caption Generation has always been a study of great interest to the researchers in the Artificial Intelligence department. The complete system is a combination of three models which optimizes the whole procedure of caption description from an image. The models are:

1. Feature Extraction Model -This model is primarily responsible for acquiring features from an image for training.
2. Encoder Model -The encoder model is primarily responsible for processing the captions of each image fed while training. The output of the encoder model is again vectors of size 1\*256 which would again be an input to the decoder sequences.
3. Decoder model-The decoder model is basically the model which concatenates both the feature extraction model and encoder model and produces the required output which is the predicted word given an image and the sentence generated till that point of time.

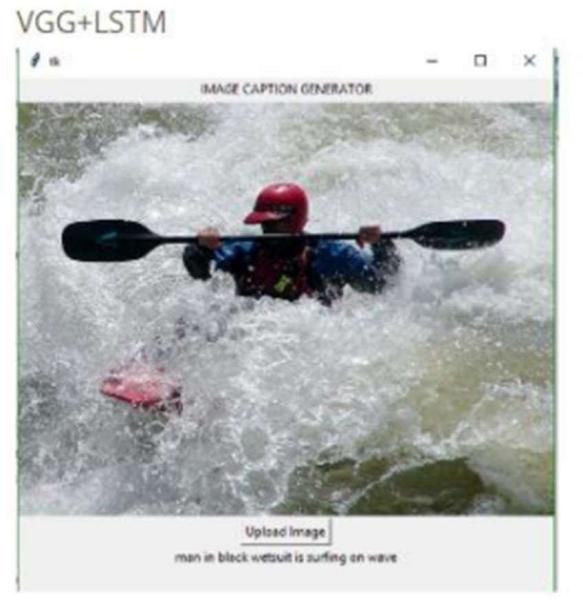
It has been tested various images with both the methods ie VGG+LSTM and VGG+GRU. The

training of the models was done on Google Colab which provides the 1xTesla K80 GPU with 12GB GDDR5 VRAM and took approximately 13 minutes per epoch for LSTM and 10 minutes per epoch for GRU. This happens

due to the lesser amount of operations occurring in GRU than LSTM. While the loss calculated for LSTM was less than GRU, the user can prefer any model according to his need, either with maximum accuracy or one which takes lesser time to process.

ANALYSIS:

some example images on which the testing was done. We have tested various images with both the methods ie VGG+LSTM and VGG+GRU.



The training of the models was done on Google Colab which provides the 1xTesla K80 GPU with 12GB GDDR5 VRAM and took approximately 13 minutes per epoch for LSTM and 10 minutes per epoch for GRU.

GRUs generally train faster on less training data than LSTMs and are simpler and easy to modify .



**Future Scope:**

The main implication of image captioning is automating the job of some person who interprets the image (in many different fields).

* 1. Probably, It will be useful in cases/fields where text is most used and with the use of this, you can infer/generate text from images. As in, use the information directly from any image in a textual format automatically.
  2. We have many NLP applications right now, which extract

insights/summary from a given text data or an essay etc. The same benefits can be obtained by people who would benefit from automated insights from pictures.

* 1. A slightly long-term use case would be, explaining what happens in a video,

frame by frame.

* 1. It is very helpful for visually impaired people. Lots of applications can be developed in that space.
  2. Social Media: Platforms like Facebook, Instagram, twitter etc can infer directly from the image, where you are (beach, cafe etc), what you wear (colour) and more importantly what you are doing also (in a way). See an example to understand it better.
  3. SkinVision: Lets you confirm whether a skin condition can be skin cancer or not.
  4. Google Photos: Classify your photo into Mountains, sea etc.

This model is not perfect and may generate incorrect captions sometimes. In the next phase, we will be developing models which will use Inceptionv3 instead of VGG as the feature extractor. Then we will be comparing the 4 models thus obtained i.e. VGG+GRU, VGG+LSTM, Inceptionv3+GRU, and Inceptionv3+LSTM . This will further help us analyze the influence of the CNN component over the entire network. Our model is trained on the Flickr 8K dataset which is relatively small with less variety of images. We will be training our model on the Flickr30K and MSCOCO datasets which will help us to make better predictions. Other optimizations include

tweaking the hyperparameters like batch size, number of epochs, learning rate etc and understanding the effect of each one of them on our model. So we venture out, explore in literature and surely will find a lot of problem are open and yet to be solved.

### References:

1. C. Wu, S. Yuan, H. Cao, Y. Wei and

L. Wang, "Hierarchical Attention-Based Fusion for Image Caption With Multi- Grained Rewards," in *IEEE Access*, vol. 8, pp. 57943-57951, 2020, doi: 10.1109/ACCESS.2020.2981513.

1. J. Gu, S. Joty, J. Cai, H. Zhao, X. Yang and G. Wang, "Unpaired Image Captioning via Scene Graph Alignments," 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 10322-10331, doi: 10.1109/ICCV.2019.01042.
2. Z. Zhu, Z. Xue and Z. Yuan, "Topic- Guided Attention for Image Captioning," *2018 25th IEEE International Conference on Image Processing (ICIP)*, 2018, pp. 2615-2619, doi: 10.1109/ICIP.2018.8451083.
3. TY - JOUR A2 - Zhang, Yin AU - Chu, Yan AU - Yue, Xiao AU - Yu, LeiAU - Sergei, Mikhailov AU - Wang, ZhengkuiPY - 2020 DA - 2020/10/21TI

- Automatic Image Captioning Based on ResNet50 and LSTM with Soft AttentionSP - 8909458 VL - 2020 SN -

15308669URhttps://doi.org/10.1155/20 20/8909458 DO -

10.1155/2020/8909458

1. A. Farhadi, M. Hejrati, M. A. Sadeghi et al., “Every picture tells a story: generating sentences from images,” in Computer Vision – ECCV 2010, K. Daniilidis, P. Maragos, and N. Paragios, Eds., pp. 15–29, Springer, 2010.
2. A. Graves, Generating sequences with recurrent neural networks, University of Toronto, 2013.
3. A. Radford, L. Metz, and S. Chintala, Unsupervised representation learning with deep convolutional generative adversarial networks, ICLR, 2016.
4. Y. Zhang, R. Gravina, H. Lu, M. Villari, and G. Fortino, “PEA: Parallel electrocardiogram-based authentication for smart healthcare systems,” Journal of Network and Computer Applications, vol. 117, pp. 10–16, 2018.
5. M. D. Z. Hossain, F. Sohel, M. F. Shiratuddin, and H. Laga, “A comprehensive survey of deep learning for image captioning,” ACM Computing Surveys, vol. 51, no. 6, pp. 1–36, 2018.
6. Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan, “Show and tell: Lessons learned from the 2015 MSCOCO image captioning challenge,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 39, no. 4, pp.652–663, 2017.
7. Hanwang Zhang, Zawlin Kyaw, Shih- Fu Chang, and Tat-Seng Chua, “Visual translation embedding network for visual relation detection,” in 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR2017, Honolulu, HI, USA, July 21-26, 2017, 2017, pp. 3107–3115.
8. Andrej Karpathy and Li Fei-Fei, “Deep visual-semantic alignments for generating image descriptions,” IEEE

Trans. Pattern Anal. Mach. Intell., vol. 39, no. 4, pp. 664–676, 2017.

1. English-speaking world, 2019. [Online; accessed 22-March-2019]. 1
2. P. Anderson, B. Fernando, M. Johnson, and S. Gould. Spice: Semantic propositional image caption evaluation.

In ECCV, 2016. 3, 6

1. P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In CVPR, 2018. 2
2. M. Artetxe, G. Labaka, E. Agirre, and K. Cho. Unsupervised neural machine translation. In ICLR, 2018.1

[17].Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition", *CoRR*, vol. abs/1409.1556, 2014.