IJNRD.ORG ISSN : 2456-4184

International Research Journal
IJNRD
Research Through Innovation

INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

GENERATING IMAGE CAPTIONS BASED ON DEEP NEURAL NETWORKS

[1]T.Sandhya, [2]Dr.Kondapalli Venkata Ramana
[1]M.Tech, [2]Associate Professor
[1,2]Dept of CS&SE, AUCE, AU, Visakhapatnam

ABSTRACT

In this project, we leverage the power of deep learning, specifically employing Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks to develop an image caption generator. This innovative endeavor capitalizes on the vast datasets and computational capabilities available in the realm of deep learning. Our goal is to create a system that can comprehend the content of an image and articulate it in English. This survey paper delves into the fundamental concepts of image captioning and outlines common methodologies. Key tools in our arsenal include the Keras library, numpy, and Jupyter notebooks. Additionally, we explore the utilization of the Flickr dataset and CNNs for image classification within the context of our project.

KEYWORDS: Deep learning techniques, Generate captions, Concepts of image captioning.

1. INTRODUCTION

Every day, we encounter numerous images across diverse sources, including the internet, news articles, documents, diagrams, and advertisements. While these images often lack detailed descriptions, humans can typically understand them intuitively. However, for machines to provide automatic image captions, they require some form of interpretation. Image captioning holds significant importance, as it can expedite and enhance the accuracy of image searches and indexing across the internet. Researchers have recognized that merely providing object names, without human-like descriptions, falls short in making a meaningful impact. The challenge lies in enabling machines to think, speak, and act more like humans in generating natural language descriptions. Image captioning finds applications in various domains such as biomedicine, commerce, web search, and military, while social media platforms like Instagram and Facebook can benefit from automatic caption generation based on images.

1.1 Motivation

Generating image captions is a crucial task at the intersection of Computer Vision and Natural Language Processing, aiming to emulate the human ability of describing images using AI. The primary challenge lies in understanding the relationships between objects within the image and conveying them in natural language, such as English. Traditionally, computer systems relied on predefined templates for generating image descriptions, limiting the diversity and richness of the generated text. However, recent advancements in neural networks have overcome this limitation, with state-of-the-art models using neural networks to input images and predict the next word in the output sentence, providing more versatile and lexically varied image captions.

1.2 Image Captioning

Process: - Image captioning is a prominent field in Artificial Intelligence, encompassing both Natural Language Processing and Computer Vision to generate textual descriptions for images. This multidisciplinary area focuses on comprehending images by detecting objects, recognizing their properties, understanding the scene or location, and deciphering object interactions. To produce coherent sentences, a combination of syntactic and semantic language understanding is essential. Effective image understanding heavily relies on extracting image features, which have applications in automatic image indexing, crucial for Content-Based Image Retrieval (CBIR) in various domains such as biomedicine, commerce, the military, education, digital libraries, and web search.

Additionally, popular social media platforms like Facebook and Twitter can automatically generate image descriptions, encompassing details about the location, attire, and activities depicted in the images.

Techniques: - The techniques for image captioning can be broadly categorized into two main approaches: traditional machine learning and deep machine learning. Traditional machine learning relies on handcrafted features like Local Binary Patterns (LBP), Scale-Invariant Feature Transform (SIFT), and Histogram of Oriented Gradients (HOG), which are extracted from input data and then used in classifiers like Support Vector Machines (SVM). However, these features are task-specific and impractical for handling diverse and complex real-world data such as images and videos. In contrast, deep machine learning techniques automatically learn features from training data, making them capable of handling diverse datasets. Convolutional Neural Networks (CNNs) are commonly used for feature learning, often followed by Recurrent Neural Networks (RNNs) or Long Short-Term Memory Networks (LSTMs) for caption generation. Deep learning algorithms excel at addressing the complexities and challenges of image captioning.

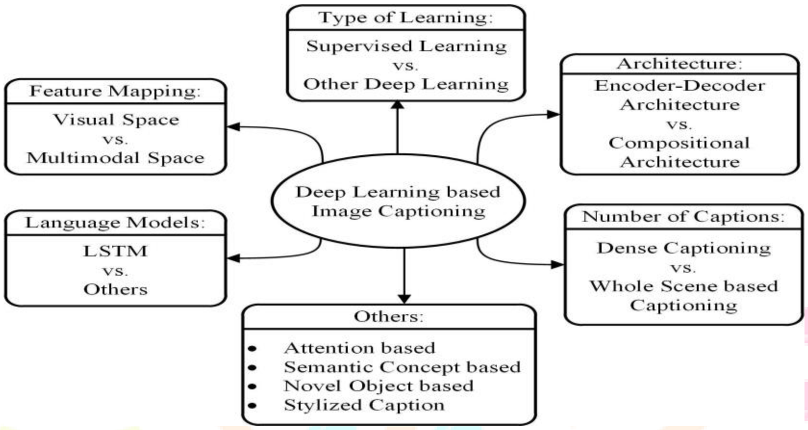


Figure.1. An overall taxonomy of deep learning-based image captioning.

2. RELATED WORK

Describing the content of images automatically is a key challenge in artificial intelligence, bridging the fields of computer vision and natural language processing. In the past, early approaches initially extracted annotations (such as nouns and adjectives) from images (Sermanet et al., 2013; Russakovsky et al., 2015), and then constructed sentences based on these annotations (Gupta and Mannem). Donahue et al. (Donahue et al.) introduced a recurrent convolutional architecture designed for large-scale visual learning, showcasing its value in three distinct tasks: video recognition, image description, and video description. These models integrated long-term dependencies into their network updates and were trainable end-to-end. However, they struggled with interpreting intermediate results.

The LRCN method was later extended to generate text from videos (Venugopalan et al.). In contrast to LRCN's single architecture for multiple tasks, Vinyals et al. (Vinyals et al.) proposed the Neural Image Caption (NIC) model, specifically designed for generating image captions. This model combines GoogLeNet with a single layer of LSTM and is trained to maximize the likelihood of generating the target description sentence from training images. NIC's performance was evaluated both qualitatively and quantitatively, and it secured the top position in the MS COCO Captioning Challenge (2015), which was assessed by human judges.

Comparing LRCN and NIC, we identify three key differences that may explain performance variations. First, NIC employs GoogLeNet, while LRCN relies on VGGNet. Second, NIC feeds visual features only into the first LSTM unit, while LRCN supplies visual features to every LSTM unit. Third, NIC features a simpler RNN architecture (single-layer LSTM) compared to LRCN (two factored LSTM layers). We confirmed that the mathematical models underlying LRCN and NIC are the same for image captioning, with performance differences arising from implementation choices. LRCN faces the challenge of balancing simplicity and generality, as it was designed for three distinct tasks.

In contrast to end-to-end learning, Fang et al. (Fang et al.) introduced a method based on visual concepts. They began by training visual detectors for commonly occurring caption words (nouns, verbs, adjectives) using multiple-instance learning. Subsequently, they trained a language model using a substantial dataset of over 400,000 image descriptions to capture word usage statistics. Finally, they ranked caption candidates using sentence-level features and a deep multi-modal similarity model. Remarkably, their captions equaled or surpassed human-written captions 34% of the time. However, this method introduces more human-controlled parameters, making it less reproducible. It is believed that the Microsoft web application caption bot is based on this approach.

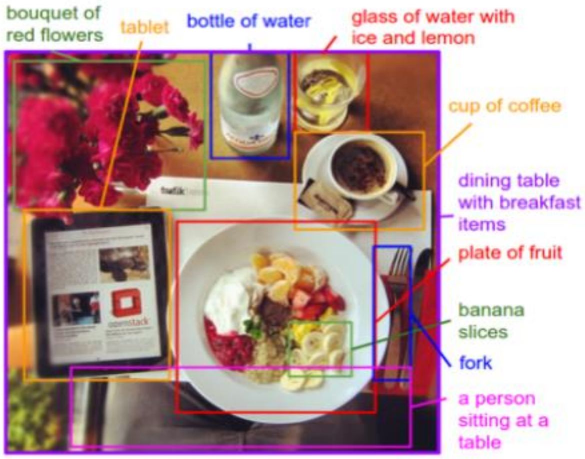

bouquet of red flowers
tablet
bottle of water
glass of water with ice and lemon
cup of coffee
dining table with breakfast items
plate of fruit
banana slices
fork
a person sitting at a table

Figure 2: The visual-semantic alignment method can generate descriptions of image regions. Figure from (Karpathy and FeiFei,)

Karpathy and Fei-Fei (Karpathy and Fei-Fei) introduced a novel approach called Visual-Semantic Alignment (VSA) to generate descriptions for distinct regions within an image, expressed as words or sentences (refer to Figure 2 for illustration). In this technique, the traditional Convolutional Neural Network (CNN) is replaced with region-based Convolutional Networks (RCNN) to ensure that the extracted visual features are closely aligned with specific regions of the image. Experimental results demonstrate that the descriptions produced by this method surpass retrieval-based benchmarks, both when applied to entire images and when evaluated on a new dataset with annotations at the region level. Notably, VSA exhibits superior performance in terms of generating more diverse and accurate descriptions compared to whole-image methods such as Long Short-Term Recurrent Convolutional Networks (LRCN) and Neural Image Captioning (NIC).

However, it is worth noting that this method comprises two separate models. Subsequently, this approach has been further extended to include dense captioning (as explored by Johnson et al. in 2016) and the development of image-based question-answering systems (as explored by Zhu et al. in 2016).

3. PROBLEM IDENTIFICATION

Despite the notable achievements of Recurrent Neural Networks (RNNs) in various applications, there persist significant challenges that demand attention. Two prominent issues commonly encountered in RNN-based systems are:

- The Vanishing Gradient Problem.
- The arduous nature of training RNNs.

Recurrent Neural Networks are a type of deep learning algorithm tailored to handle intricate computational tasks like object classification and speech recognition. They excel in scenarios where events occur sequentially, as they leverage information from prior events to comprehend subsequent ones.

In an ideal scenario, we would prefer RNNs to be deeper, thus extending their memory capacity and enhancing their capabilities. This would open possibilities for practical applications, such as stock price prediction and improved speech recognition. However, the reality is that RNNs are seldom employed in real-world scenarios due to the persistent challenge posed by the vanishing gradient problem.

3.1 The Vanishing Gradient Problem

This is one of the most significant challenges for RNNs performance. In practice, the architecture of RNNs restricts its long-term memory capabilities, which are limited to only remembering a few sequences at a time. Consequently, the memory of RNNs is only useful for shorter sequences and short time-periods.

Vanishing Gradient problem arises while training an Artificial Neural Network. This mainly occurs when the network parameters and hyperparameters are not properly set. The vanishing gradient problem restricts the memory capabilities of traditional RNNs—adding too many time-steps increases the chance of facing a gradient problem and losing information when you use backpropagation.

4. PROPOSED WORK

The main aim of this project is to get a little bit of knowledge of deep learning techniques. We use two techniques mainly CNN and LSTM for image classification.

So, to make our image caption generator model, we will be merging these architectures. It is also called a CNN-RNN model.

- CNN is used for extracting features from the image. We will use the pre-trained modelXception.
- LSTM will use the information from CNN to help generate a description of the image.

4.1 Convolutional Neural Network

A Convolutional Neural Network (ConvNet/CNN) is a Deep Learning algorithm which can take in an input image, assign importance (learnable weights and biases) to various aspects/objects in the image and be able to differentiate one from the other.The pre-processing required in a ConvNet is much lower as compared to other classification algorithms. Convolutional Neural networks are specialized deep neural networks which can process the datathat has input shape like a 2D matrix. Images are easily represented as a 2D matrix and CNN is very useful in working with images. It scans images from left to right and top to bottom to pull out important features from the image and combines the feature to classify images. It can handle the images that have been translated, rotated, scaled and changes in perspective.

4.2 Long Short Term Memory

LSTM stands for Long short term memory, they are a type of RNN (recurrent neural network) which is well suited for sequence prediction problems. Based on the previous text, we canpredict what the next word will be. It has proven itself effective from the traditional RNN by overcoming the limitations of RNN which had short term memory. LSTM can carry out relevant information throughout the processing of inputs and with a forget gate, it discards non-relevant information.

LSTMs are designed to overcome the vanishing gradient problem and allow them to retain information for longer periods compared to traditional RNNs. LSTMs can maintain a constant error, which allows them to continue learning over numerous time-steps and backpropagate through time and layers.

LSTMs use gated cells to store information outside the regular flow of the RNN. With thesecells, the network can manipulate the information in many ways, including storing informationin the cells and reading from them. The cells are individually capable of making decisions regarding the information and can execute these decisions by opening or closing the gates. The ability to retain information for a long period of time gives LSTM the edge over traditional RNNs in these tasks. The chain-like architecture of LSTM allows it to contain information for longer time periods, solving challenging tasks that traditional RNNs struggle to or simply cannot solve.

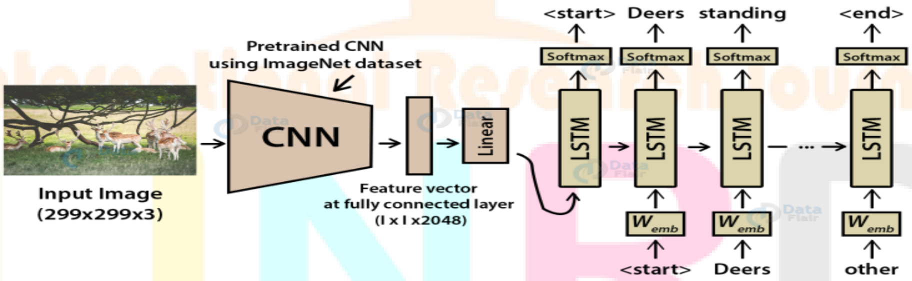


Figure. 3. Model, Image Caption Generator

The three major parts of the LSTM include:

Forget gate—removes information that is no longer necessary for the completion of the task. This step is essential to optimizing the performance of the network.

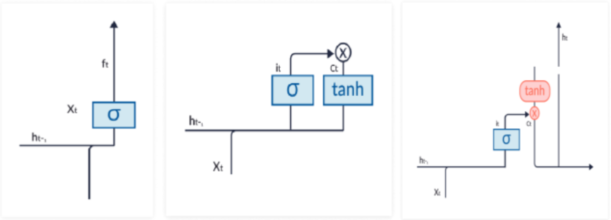


Figure.4. Forget Gate,Input Gate, Output Gate

Input gate—responsible for adding information to the cells

Output gate—selects and outputs necessary information

The CNN LSTM architecture involves using Convolutional Neural Network (CNN) layers for feature extraction on input data combined with LSTMs to support sequence prediction. This architecture was originally referred to as a Long-term Recurrent Convolutional Network or LRCN model, although we will use the more generic name “CNN LSTM” to refer to LSTMs that use a CNN as a front end in this lesson. This architecture is used for the task of generating textual descriptions of images. Key is the use of a CNN that is pre-trained on a challenging image classification task that is re-purposed as a feature extractor for the caption generating problem.

5. DATASET

Flickr8k dataset is a public benchmark dataset for image to sentence description. This dataset consists of 8000 images with five captions for each image. These images are extracted from diverse groups in Flickr website. Each caption provides a clear description of entities and events present in the image. The dataset depicts a variety of events and scenarios and doesn‟t include images containing well-known people and places which makes the dataset more generic. The dataset has 6000 images in training dataset, 1000 images in development dataset and 1000 images in test dataset. Features of the dataset making it suitable for this project are: • Multiple captions mapped for a single image makes the model generic and avoids overfitting of the model.

- Diverse category of training images can make the image captioning model to work for multiple categories of images and hence can make the model more robust.

Image Data Preparation

The image should be converted to suitable features so that they can be trained into a deeplearning model. Feature extraction is a mandatory step to train any image in deep learning model.The features are extracted using Convolutional Neural Network (CNN) with Visual Geometry Group (VGG-16) model. This model also won ImageNet Large Scale Visual Recognition Challenge in 2015 to classify the images into one among the 1000 classes given in the challenge. Hence, this model is ideal to use for this project as image captioning requires identification of images.

In VGG-16, there are 16 weight layers in the network and the deeper number of layers help in better feature extraction from images. The VGG-16 network uses 3*3 convolutional layersmaking its architecture simple and uses max pooling layer in between to reduce volume size of

the image. The last layer of the image which predicts the classification is removed and the internal representation of image just before classification is returned as feature. The dimension ofthe input image should be 224*224 and this model extracts features of the image and returns a 1- dimensional 4096 element vectors.

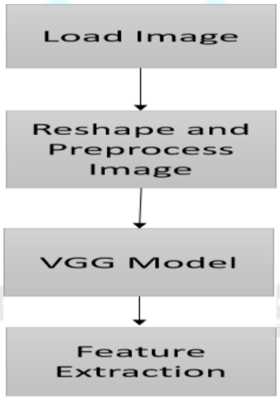


Figure 5: Feature Extraction in images using VGG

Caption Data Preparation

Flickr8k dataset contains multiple descriptions described for a single image. In the

data preparation phase, each image id is taken as key and its corresponding captions are stored as values in a dictionary.

Data Cleaning

In order to make the text dataset work in machine learning or deep learning models, raw text should be converted to a usable format. The following text cleaning steps are done before using it for the project: • Removal of punctuations. • Removal of numbers. • Removal of single length words. • Conversion of uppercase to lowercase characters. Stop words are not removed from the text data as it will hinder the generation of a grammatically complete caption which is needed forthis project. Table 1 shows samples of captions after data cleaning.

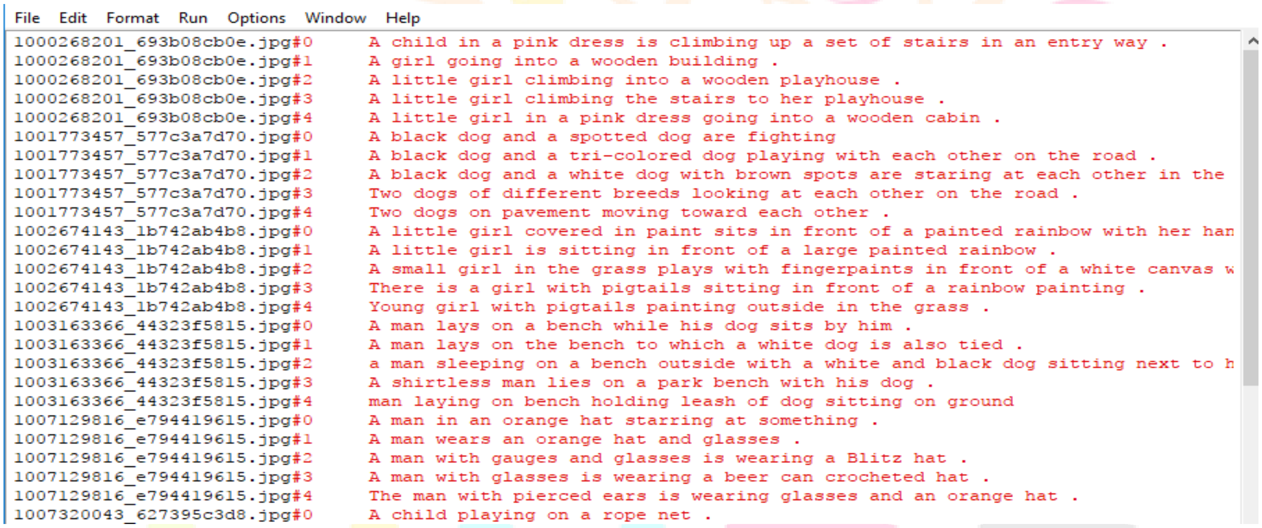
Original Captions	Captions after Data cleaning
Two people are at the edge of a lake, facing the water and the city skyline.	people are at the edge of lake facingthe water and the city skyline
A little girl rides in a child 's swing.	little girl rides in child swing
Two boys posing in blue shirts and khaki shorts.	Two boys posing in blue shirts and khaki shorts

Table 1: Data cleaning of captions

6. IMPLEMENTATION

Data Cleaning

The main text file which contains all image captions is **Flickr8k.token** in our **Flickr_8k_text** folder.



```
1000268201_693b08cb0e.jpg#0	A child in a pink dress is climbing up a set of stairs in an entry way .
1000268201_693b08cb0e.jpg#1	A girl going into a wooden building .
1000268201_693b08cb0e.jpg#2	A little girl climbing into a wooden playhouse .
1000268201_693b08cb0e.jpg#3	A little girl climbing the stairs to her playhouse .
1000268201_693b08cb0e.jpg#4	A little girl in a pink dress going into a wooden cabin .
1001773457_577c3a7d70.jpg#0	A black dog and a spotted dog are fighting
1001773457_577c3a7d70.jpg#1	A black dog and a tri-colored dog playing with each other on the road .
1001773457_577c3a7d70.jpg#2	A black dog and a white dog with brown spots are staring at each other in the
1001773457_577c3a7d70.jpg#3	Two dogs of different breeds looking at each other on the road .
1001773457_577c3a7d70.jpg#4	Two dogs on pavement moving toward each other .
1002674143_1b742ab4b8.jpg#0	A little girl covered in paint sits in front of a painted rainbow with her han
1002674143_1b742ab4b8.jpg#1	A little girl is sitting in front of a large painted rainbow .
1002674143_1b742ab4b8.jpg#2	A small girl in the grass plays with fingerpaints in front of a white canvas w
1002674143_1b742ab4b8.jpg#3	There is a girl with pigtails sitting in front of a rainbow painting .
1002674143_1b742ab4b8.jpg#4	Young girl with pigtails painting outside in the grass .
1003163366_44323f5815.jpg#0	A man lays on a bench while his dog sits by him .
1003163366_44323f5815.jpg#1	A man lays on the bench to which a white dog is also tied .
1003163366_44323f5815.jpg#2	a man sleeping on a bench outside with a white and black dog sitting next to h
1003163366_44323f5815.jpg#3	A shirtless man lies on a park bench with his dog .
1003163366_44323f5815.jpg#4	man laying on bench holding leash of dog sitting on ground
1007129816_e794419615.jpg#0	A man in an orange hat starring at something .
1007129816_e794419615.jpg#1	A man wears an orange hat and glasses .
1007129816_e794419615.jpg#2	A man with gauges and glasses is wearing a Blitz hat .
1007129816_e794419615.jpg#3	A man with glasses is wearing a beer can crocheted hat .
1007129816_e794419615.jpg#4	The man with pierced ears is wearing glasses and an orange hat .
1007320043_627395c3d8.jpg#0	A child playing on a rope net .
```

Figure.6. Flicker DataSet text format

The format of our file is image and caption separated by a new line (“\n”). Each image has 5 captions and we can see that #(0 to 5)number is assigned for each caption.We will define 5 functions:

The 'load_doc' function loads a document file and reads its contents into a string. The 'all_img_captions' function creates a dictionary mapping images to lists of five captions. The 'cleaning_text' function cleans the descriptions by removing punctuation, converting text to lowercase, and eliminating words containing numbers. The 'text_vocabulary' function extracts unique words to create a vocabulary from the descriptions. Lastly, the 'save_descriptions' function saves the preprocessed descriptions into a 'descriptions.txt' file.

Extracting The Feature Vector from All Images

This technique is also called transfer learning, we don’t have to do everything on our own, we use the pre-trained model that have been already trained on large datasets and extract the features from these models and use them for our tasks. We are using the Xception model which has been trained on imagenet dataset that had 1000 different classes to classify. We can directly importthis model from the keras.applications . Make sure you are connected to the internet as the weights get automatically downloaded. Since the Xception model was originally built for imagenet, we will do little changes for integrating with our model. One thing to notice is that the Xception model takes 299*299*3 image size as input. We will remove the last classification layer and get the 2048 feature vector. The function **extract_features()** will extract features for all images and we will map image names with their respective feature array. Then we will dump the features dictionary into a “features.p” pickle file. This process can take a lot of time depending on your system. I am using an GPU for training purpose so it took me around 7 minutes for performing this task. However, if you areusing CPU then this process might take 1-2 hours.

Tokenizing The Vocabulary

Computers don‟t understands English words, for computers, we will have to represent them with numbers. So, we will map each word of the vocabulary with a unique index value. Keras library provides us with the tokenizer function that we will use to create tokens from our vocabulary andsave them to a **“tokenizer.p”** pickle file.

Our vocabulary contains 7577 words. We calculate the maximum length of the descriptions. This is important for deciding the model structure parameters. Max_length of description is 32.

Create Data generator

Let us first see how the input and output of our model will look like. To make this task into a supervised learning task, we must provide input and output to the model for training. We must train our model on 6000 images and each image will contain 2048 length feature vector and caption is also represented as numbers. This amount of data for 6000 images is not possible to hold into memory so we will be using a generator method that will yield batches. The generator will yield the input and output sequence.

For example:

The input to our model is [x1, x2] and the output will be y, where x1 is the 2048 feature vector of that image, x2 is the input text sequence and y is the output text sequence that the model has to predict.

x1(feature vector)	x2(Text sequence)	y(word to predict)
feature	start,	two
feature	start, two	dogs
feature	start, two, dogs	drink
feature	start, two, dogs, drink	water
feature	start, two, dogs, drink, water	end

Table 2. Word Prediction Generation Step By Step

Defining the CNN-RNN model

To define the structure of the model, we will be using the Keras Model from Functional API. It will consist of three major parts:

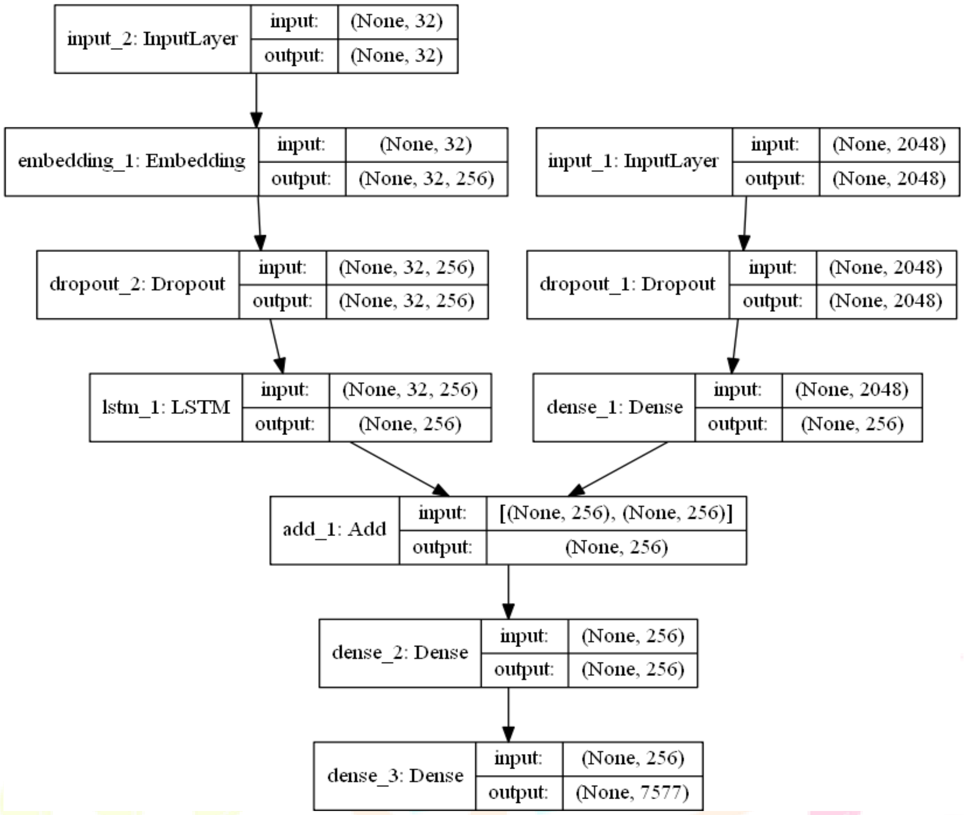
Feature Extractor – The feature extracted from the image has a size of 2048, with a dense layer, we will reduce the dimensions to 256 nodes.

Sequence Processor – An embedding layer will handle the textual input, followed by the LSTM layer.

Decoder – By merging the output from the above two layers, we will process by the dense layer to make the final prediction. The final layer will contain the number of nodes equal to our vocabulary size.

Training the model

To train the model, we will be using the 6000 training images by generating the input and output sequences in batches and fitting them to the model using model.fit_generator() method. We also save the model to our model’s folder. This will take some time depending on your system capability.

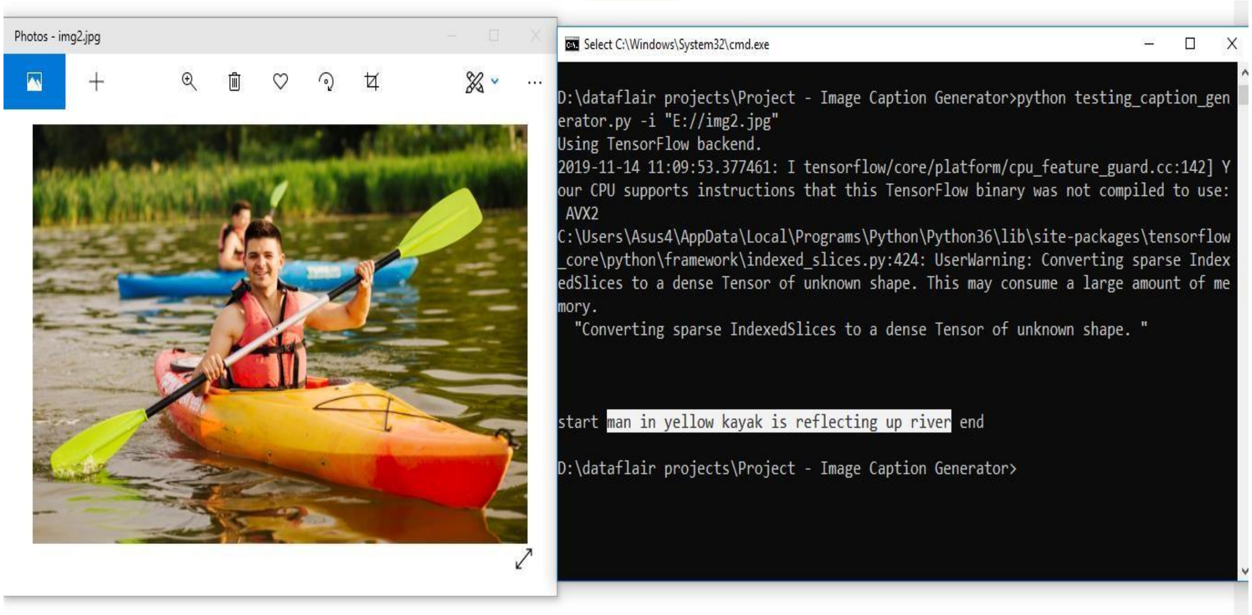


.

Figure.7. Final Model Sructure

Testing the model

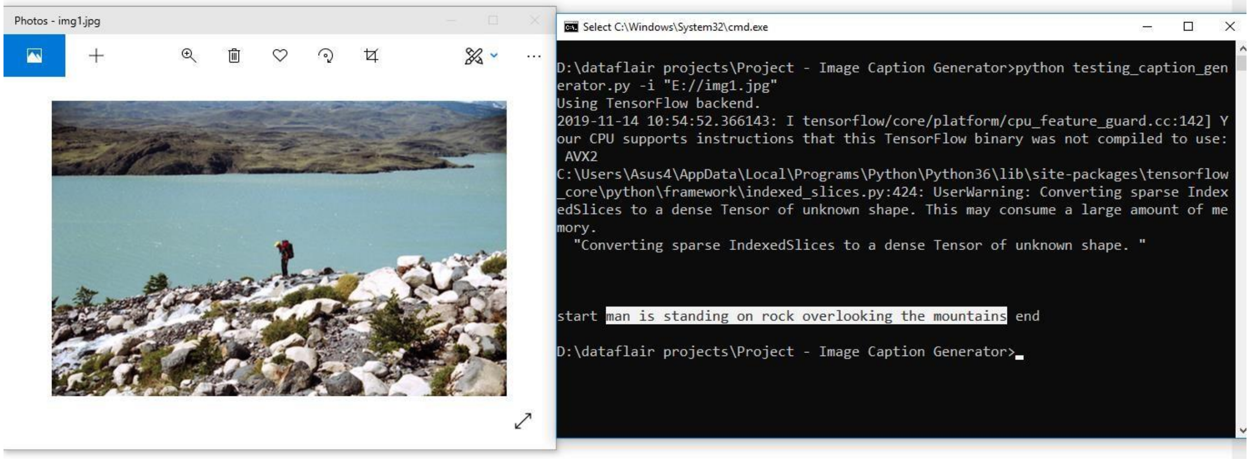
The model has been trained, now, we will make a separate file testing_caption_generator.py which will load the model and generate predictions. The predictions contain the max length of index values so we will use the same tokenizer.p pickle file to get the words from their index values.



```
D:\dataflair projects\Project - Image Caption Generator>python testing_caption_gen
erator.py -i "E://img2.jpg"
Using TensorFlow backend.
2019-11-14 11:09:53.377461: I tensorflow/core/platform/cpu_feature_guard.cc:142] Y
our CPU supports instructions that this TensorFlow binary was not compiled to use:
 AVX2
C:\Users\Asus4\AppData\Local\Programs\Python\Python36\lib\site-packages\tensorflow
_core\python\framework\indexed_slices.py:424: UserWarning: Converting sparse Index
edSlices to a dense Tensor of unknown shape. This may consume a large amount of me
mory.
  "Converting sparse IndexedSlices to a dense Tensor of unknown shape. "


start man in yellow kayak is reflecting up river end

D:\dataflair projects\Project - Image Caption Generator>
```



```
D:\dataflair projects\Project - Image Caption Generator>python testing_caption_gen
erator.py -i "E://img1.jpg"
Using TensorFlow backend.
2019-11-14 10:54:52.366143: I tensorflow/core/platform/cpu_feature_guard.cc:142] Y
our CPU supports instructions that this TensorFlow binary was not compiled to use:
 AVX2
C:\Users\Asus4\AppData\Local\Programs\Python\Python36\lib\site-packages\tensorflow
_core\python\framework\indexed_slices.py:424: UserWarning: Converting sparse Index
edSlices to a dense Tensor of unknown shape. This may consume a large amount of me
mory.
  "Converting sparse IndexedSlices to a dense Tensor of unknown shape. "


start man is standing on rock overlooking the mountains end

D:\dataflair projects\Project - Image Caption Generator>
```

Figure.8. Output Caption of Given Image

7. CONCLUSION

In this paper, we have reviewed deep learning-based image captioning methods. We have given a taxonomy of image captioning techniques, shown generic block diagram of the major groups and highlighted their pros and cons. We discussed different evaluation metrics and datasets with

their strengths and weaknesses. A brief summary of experimental results is also given. We briefly outlined potential research directions in this area. Although deep learning-based image captioning methods have achieved a remarkable progress in recent years, a robust image captioning method that is able to generate high quality captions for nearly all images is yet to be achieved. With the advent of novel deep learning network architectures, automatic image captioning will remain an active research area for some time.

We have used Flickr_8k dataset which includes nearly 8000 images, and the corresponding captions are also stored in the text file. Although deep learning -based image captioning methods have achieved a remarkable progress in recent years, a robust image captioning method that is able to generate high quality captions for nearly all images is yet to be achieved. With the advent of novel deep learning network architectures, automatic image captioning will remain an active research area for some time. The scope of image-captioning is very vast in the future as the users are increasing day by day on social media and most of them would post photos. So this project will help them to a greater extent.

Limitations

The neural image caption generator gives a useful framework for learning to map from images to human-level image captions. By training on large numbers of image-caption pairs, the model learns to capture relevant semantic information from visual features.

However, with a static image, embedding our caption generator will focus on features of our images useful for image classification and not necessarily features useful for caption generation. To improve the amount of task-relevant information contained in each feature, we can train the image embedding model (the VGG-16 network used to encode features) as a piece of the caption generation model, allowing us to fine-tune the image encoder to better fit the role of generating captions. Also, if we actually look closely at the captions generated, we notice that they are rathermundane and commonplace. Take this possible image-caption pair for instance:

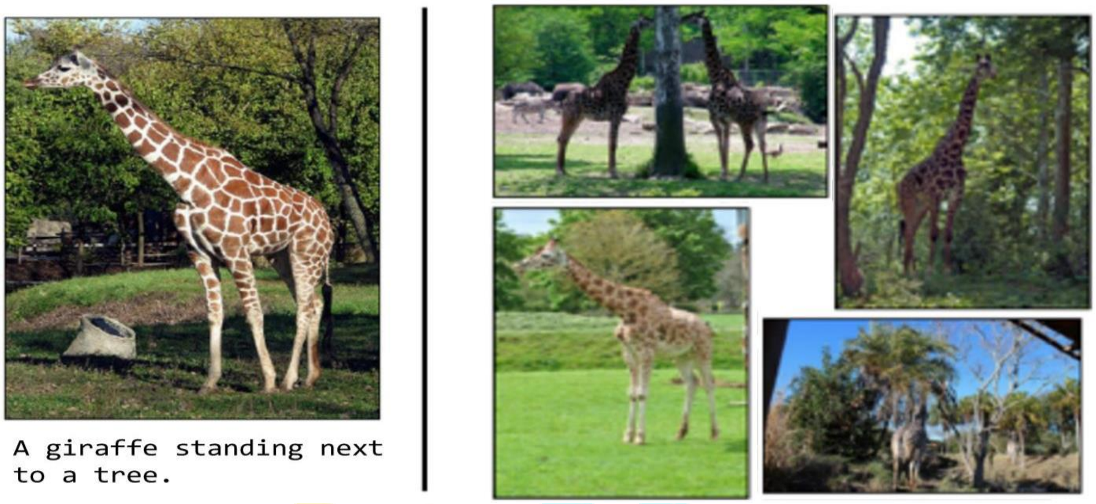


Figure.9. The above picture depicts clear limitation of the model because it rely most onthe training dataset

This is most certainly a “giraffe standing next to a tree.” However, if we look at other pictures, we will likely notice that it generates a caption of “a giraffe next to a tree” for any picture with a giraffe because giraffes in the training set often appear near trees.

Future Scope

Future work Image captioning has become an important problem in recent days due to the exponential growth of images in social media and the internet. This report discusses the various research in image retrieval used in the past and it also highlights the various techniques and methodology used in the research. As feature extraction and similarity calculation in images are challenging in this domain, there is a tremendous scope of possible research in the future.Current image retrieval systems use similarity calculation by making use of features such as color, tags, image retrieval using image captioning 54 histogram, etc. There cannot be completely accurate results as these methodologies do not depend on the context of the image. Hence, complete research in image retrieval making use of context of the images suchas image captioning will facilitate to solve this problem in the future. This project can be further enhanced in future to improve the identification of classes which has a lower precision by training it with more image captioning datasets. This methodology can also be combined with previous image retrieval methods such as histogram, shapes, etc. and can be checked if the image retrieval results get better.

8. REFERENCES

[1] Abhaya Agarwal and Alon Lavie. 2008. Meteor, m-bleu and m-ter: Evaluation metrics for high-correlation with human rankings of machine translation output. In Proceedings of the ThirdWorkshop on Statistical Machine Translation. Association for Computational Linguistics, 115–118.

[2] Ahmet Aker and Robert Gaizauskas. 2010. Generating image descriptions using dependency relational patterns. In Proceedings of the 48th annual meeting of the association for computational linguistics.Association for Computational Linguistics, 1250–1258.

[3] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. Spice: Semantic propositional image caption evaluation. In European Conference on Computer Vision. Springer, 382–398.

[4] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2017. Bottom-up and top-down attention for image captioning and vqa. arXiv preprintarXiv:1707.07998 (2017).

[5] Jyoti Aneja, Aditya Deshpande, and Alexander G Schwing. 2018. Convolutional image captioning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 5561–5570.

[6] Lisa Anne Hendricks, Subhashini Venugopalan, Marcus Rohrbach, Raymond Mooney, Kate Saenko, Trevor Darrell, Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, et al. 2016. Deep compositional captioning: Describing novel object categories without paired training data. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.

[7] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translationby jointly learning

to align and translate. In International Conference on Learning Representations (ICLR).

[8] Shuang Bai and Shan An. 2018. A Survey on Automatic Image Caption Generation. Neurocomputing. ACM Computing Surveys, Vol. 0, No. 0, Article 0. Acceptance Date: October 2018.0:30 Hossain et al.

[9] Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization, Vol. 29. 65–72.

[10] Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. 2015. Scheduled sampling for sequence prediction with recurrent neural networks. In Advances in Neural Information Processing Systems. 1171–1179.

[11] Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. 2003. A neural probabilistic language model. Journal of machine learning research 3, Feb, 1137–1155.

[12] Adam L Berger, Vincent J Della Pietra, and Stephen A Della Pietra. 1996. A maximum entropy approach to natural language processing. Computational linguistics 22, 1 (1996), 39–71.

[13] Raffaella Bernardi, Ruket Cakici, Desmond Elliott, Aykut Erdem, Erkut Erdem, Nazli Ikizler-Cinbis, Frank Keller, Adrian Muscat, Barbara Plank, et al. 2016. Automatic Description Generation from Images: A Survey of Models, Datasets, and Evaluation Measures. Journal of Artificial Intelligence Research (JAIR) 55, 409–442.

[14] David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. Journal of machine Learning research 3, Jan (2003), 993–1022.

[15] Cristian Bodnar. 2018. Text to Image Synthesis Using Generative Adversarial Networks. arXiv preprint arXiv:1805.00676.

[16] Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008.

Freebase: a collaboratively created graph database for structuring human knowledge. In Proceedings of the 2008 ACM SIGMODinternational conference on Management of data. AcM, 1247–1250.

[17] Bernhard E Boser, Isabelle M Guyon, and Vladimir N Vapnik. 1992. A training algorithm for optimal margin classifiers. In Proceedings of the fifth annual workshop on Computational learning theory. ACM, 144–152.

[18] Andreas Bulling, Jamie A Ward, Hans Gellersen, and Gerhard Troster. 2011. Eye movement analysis for activity recognition using electrooculography. IEEE transactions on pattern analysis and machine intelligence 33, 4 (2011), 741–753.

[19] Vladimir Bychkovsky, Sylvain Paris, Eric Chan, and Frédo Durand. 2011. Learning photographic global tonal adjustment with a database of input/output image pairs. In Computer Vision and Pattern Recognition (CVPR), 2011