DETECTING AND CAPTIONING IMAGES USING DEEP NEURAL NETWORKS AND FLASK

Samiksha Ashtikar[1], Aishwarya Rastogi[2] , Khan Monazzam[3], Salman Akhtar[4]
[1,2,3,4]Scholars, Computer Engineering, SKN Sinhgad Institute of Technology & Science, Lonavala

ABSTRACT:

Captioning images automatically is one of the heart of the human visual system. There are various advantages if there is an application which automatically caption the scenes surrounded by them and revert back the caption as a plain message. In this paper, we present a model based on CNN-LSTM neural networks which automatically detects the objects in the images and generates descriptions for the images. It uses various pre-trained models to perform the task of detecting objects and uses CNN and LSTM to generate the captions. It uses Transfer Learning based pre-trained models for the task of object Detection. This model can perform two operations. The first one is to detect objects in the image using Convolutional Neural Networks and the other is to caption the images using RNN based LSTM(Long Short Term Memory). Interface of the model is developed using flask rest API, which is a web development framework of python. The main use case of this project is to help visually impaired to understand the surrounding environment and act according to that.

1. INTRODUCTION

Caption generation is one of the interesting and focussed areas of Artificial Intelligence which has many challenges to pass on. Caption generation involves various complex scenarios starting from picking the dataset, training the model, validating the model, creating pre-trained models to test the images , detecting the images and finally generating the captions. There are various datasets available as open source to train the model like flickr8k, flickr30k and MSCOCO. Every dataset is contained with times to training and validation images to train the model.

Our model uses two different neural networks to generate the captions. The first neural network is Convolutional Neural Network(CNN), which is used to train the images as well as to detect the objects in the image with the help of various pre-trained models like VGG, Inception or YOLO. The second neural network used is Recurrent Neural Network(RNN) based Long Short Term Memory(LSTM), which is used to generate captions from the generated object keywords.

As, there is lot of data involved to train and validate the model, generalized machine learning algorithms will not work. Deep Learning has been evolved from the recent solve the data constraints on Machine Learning algorithms. GPU based computing is required to perform the Deep Learning tasks more effectively.

2. LITERATURE SURVEY

The problem of image captioning is a complex and widely interested research topic since the evolution of deep learning. There are many proposed solutions for this problem which are replacing the previous solutions every single day. In [1] Karpathy proposed a system which uses multimodel neural networks to generate novel descriptions of the image by providing suitable descriptions for the image.

In [2], Deng proposed a model which uses a database called ImageNet which is build using the core called WordNet.

This model uses ImageNet to generate sentence descriptions from the imag.Kelvin at el [3] proposed an attention based model, which generate captions of the images based on the region of interest. It generates the captions based on the region the image is surrounded.

In [4] , Yang proposed a multimodal recurrent neural network based model, which generates the descriptions of the image by detecting the objects and converting them to sentences, . which is almost similar to human visual system. In [5], Aneja proposed a convolutional neural network based modal to generate descriptions from the image after the rigorous training given to the model.

In [6], Pan proposed a multiple neural network model, which is expermimented with large sets of datasets to generate the accurate sentence descriptions from the image.

In [7], Vinyals proposed a model that uses Natural Language Processing and Computer Vision to detect the objects in the image and generate captions based on word processing and keyword retrieval techniques.

3. DATASET

In performing the task of image captioning, we have used flickr8k dataset from flickr.com website. It consists of a total of 8092 images. Those 8092 images were splitted into 6000 training images and 1000 each for development and testing purpose. It consists of daily life images with features covering many objects. The images were of high clarity with good resolution and were easily recognizable for the model to get trained. It is an open source dataset, which is available freely on the internet.

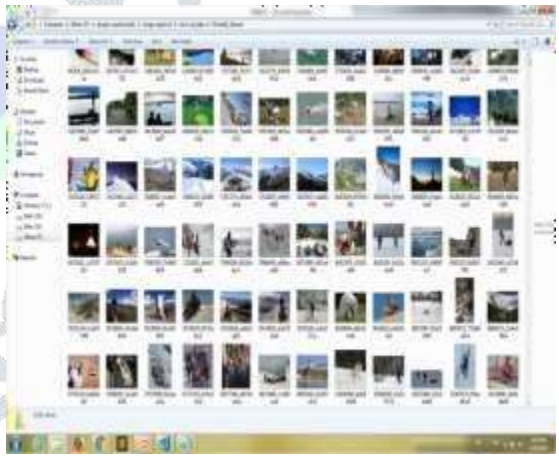

Figure 1: Flickr8k datas

4. PHASES OF MODEL:

4.1. Creating a pre-trained model using Transfer Learning:

Transfer Learning is used to create and use pre-trained models in solving complex machine learning problems. It is the way of preserving knowledge we gained by

solving a problem and use it later to solve another complex problem.

In this model, we are using a pre-trained model with Transfer learning to make our model learn things from the existing knowledge.

4.2. Object Detection:

Objects in the images were detected in this phase. It uses Convolutional Neural Network(CNN) to extract the features from

4.3. Probabilistic NLP model:

The objects detected in the image were sent through this NLP(Natural Language Processing) based probabilistic model, which removes the unnecessary features from the image. It processes only the features which are relevant and meaningful with context of the image and ignore the odd ones which are irrelevant. It also removes the stop words which are repeated and have the same meaning.

the image. We will use pre-trained model like Inception V4 or VGG 16 which is a Convolutional Neural Network for the task of object detection.

Each and every object in the image is detected in this phase.The objects were also marked with it's name.

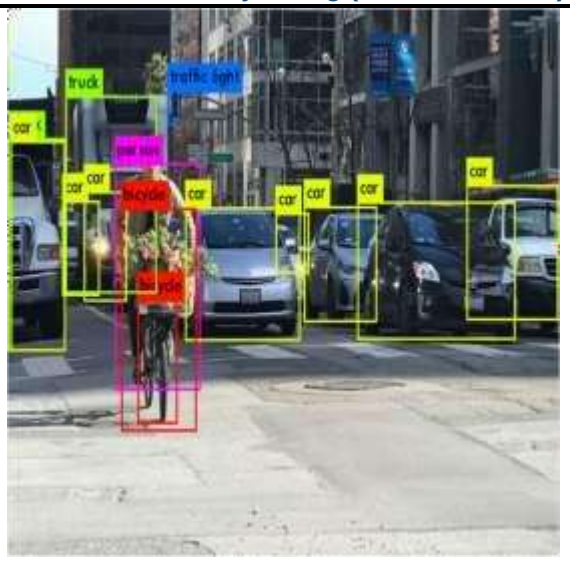

Figure 2 : Object Detection

4.4. Caption Generation: This phase combines both object detection phase and probabilistic model phase to generate the captions for the image. The output of the above phases is given to Long Short Term Memory(LSTM), which is a special type of Recurrent Neural Network(RNN) to generate the captions. LSTMs are used to hold long term dependencies. It allows RNN continue to learn over multiple steps by maintaining a consistent error by preserving error that can be backpropagated through layers and time.Captions were produced in this phase.

4.5. Ranking based Caption Retrieval:

Different layers in LSTMs generates multiple captions from the image. In this phase, the captions generated from the top layers of the LSTMs were ranked based on the captions which were supported by more numbers of LSTM layers. The caption which gets the top rank will be treated as the final caption. Different layers of LSTM generate it's own captions. The caption

which is supported by more number of LSTM's will be taken as the final caption.

4.6. Deployment to Web Server: The final phase of the project is to deploy the caption generation model in the form of web application. We are using Flask Rest API, which is a web framework of Python to deploy the working model. Flask is one of the popular Python Web development framework to develop and deploy models into web applications. We also use html, css and bootstrap to design the interface of the web application.

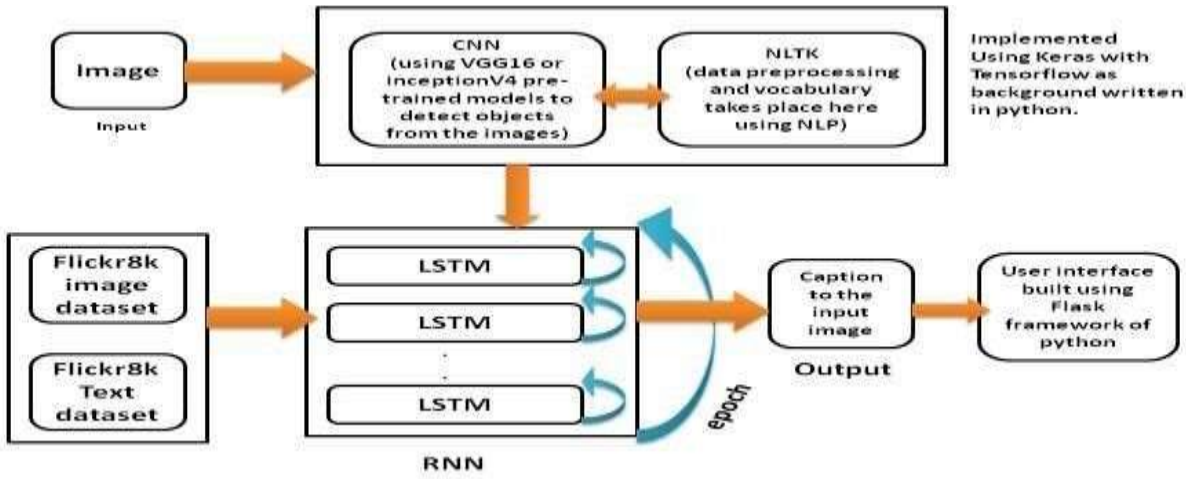

Figure 3: Architecture of the Model

6. TRAINING PHASE

In Training phase, we provide pair of image dataset as well as the captions of these images to the model. VGG model, which is pre-trained can able to detect all the possible objects present in the image. While LSTM portion is used to

predict each and every word from the image after observing the image. To each and every caption, we add starting and ending symbols to recognize the sentence. If any stop word is approached in the sentence, it stops sentence formation and marks as the end of the string. We need to calculate Loss function by using the below formula. In the formula, 'I' represents the input image and 'S' represents the generated caption. In the process of training , we need to minimize the loss function.

$$L(I,S) = -\sum_{t=1}^{N} \log p_t(S_t)$$

After training the dataset with the RNN based Long Short Term Memory(LSTM), the descriptions of the trained images were stored in a separate text file as shown in the below figure.

Figure 4: Captions of the trained images

The pseudo code form implementing the LSTM is as follows:

$cf_t = \sigma_1(Wc\ f \bullet [Ot{-}1, xt] + bc\ f)$ (1)

$I_t = \sigma_2(WI \bullet [O_{t-1}, x_t] + b_I)$ (2)

$S_t = \tanh(W_S \bullet [O_{t-1}, x_t] + b_S)$ (3)

$S_t = cf_t \times St_1 + I_t \times S_{t-1}$ (4)

7. IMPLEMENTATION AND RESULTS

Python is used to implement this model. It uses Scipy environment. Scipy is a machine learning library of Python. It uses Keras environment. Keras is a deep learning library written in Python. This model uses Tensorflow as a backend, which is popular Deep Learning framework of Python. Convolutional Neural Network and Long Short Term Memory are two neural networks used to implement this model. Transfer Learning is used to create and use pre-trained models to deal with complex object detection tasks. Flickr8k dataset is trained with the model, which generates sentence based captions using CNN to detect the objects and LSTM to generate the captions from the input image. The interface is used to deploy our model with the help of Flask Rest API of python.

Figure 5: Input 1 to the model

Consider the above image as the input to the model, the output to the given input is given below.

Figure 6: Output to the Input 1

The evaluation diagram is shown below:

8. EVALUATION

We have evaluated our model with various existing model and our model performed better than all the models in every case. We have tested with 100 images of four different varieties includes people faces, location specific images, animal images and other objects. Our model outperformed other models in all the four cases.

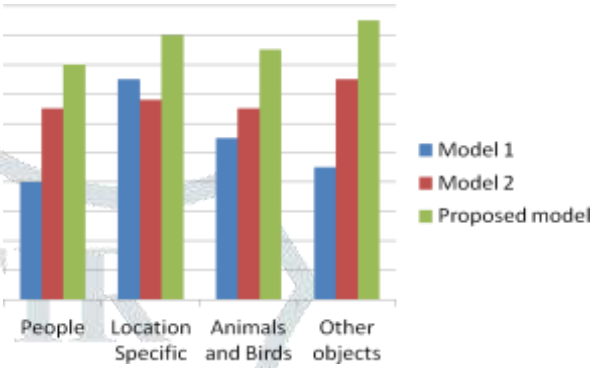

Figure 9: Evaluation of the proposed model

9. ADVANTAGES

There are various advantages of Image captiog in multiple disciplines.

captioning live video frames cannot be possible with the general CPUs. Video captioning is a popular research area in which it is going to change the lifestyle of the people with the use cases being widely usable in almost every domain. It automates the major tasks like video surveillance and other security tasks.

.

10. FUTURE WORK

We are going to extend our work in the next higher level by enhancing our model to generate captions even for the live video frame. Our present model generates captions only for the image, which itself a complex task and captioning live video frames is much complex to create. This is completely GPU based and

11. CONCLUSION

Image captioning has many advantages in almost every complex area of Artificial Intelligence. The main use case of our model is to help visually impaired to understand the environment and made them easy to act according to the environment. As, this is a complex task to do, with the help of pre trained models and powerful

deep learning frameworks like Tensorflow and Keras, we made it possible. This is completely a Deep Learning project, which makes use of multiple Neural Networks like Convolutional Neural Network and Long Short Term Memory to detect objects and captioning the images. To deploy our model as a web application, we have used Flask, which is a powerful Python's web framework.

11. REFERENCES

[1] Andrej Karpathy, Li Fei-Fei, Deep VisualSemantic Alignments for Generating Image Descriptions,[Online] Available: https://cs.stanford.edu/people/karpathy/cvpr2015.pdf.

[2] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li and Li Fei-Fei, ImageNet: A Large-Scale Hierarchical Image Database.

[3] Kelvin Xu, Jimmy Lei Ba, Ryan Kiros, Cho, Aaron Courville, Ruslan Salakhutdinov, Richard S. Zemel, Yoshua Bengio, Show, Attend and Tell: Neural Image Caption Generation with Visual Attention, [Online] Available: https://arxiv.org/pdf/1502.03044.pdf [9] M. Hodosh, P. Young and J. Hockenmaier (2013) "Framing Image Description as a Ranking Task: Data, Models and Evaluation Metrics", Journal of Artificial Intelligence Research, Volume 47, pages 853-899.

[4] Zhongliang Yang, Yu-Jin Zhang, Sadaqat ur Rehman, Yongfeng Huang, Image Captioning with Object Detection and Localization, [Online] Available: https://arxiv.org/ftp/arxiv/papers/1706/1706.02430.pdf.

[5] Jyoti Aneja, Aditya Deshpande, Alexander Schwing, Convolutional Image Captioning.

[6] Jia-Yu Pan, Hyung-Jeong Yang, Pinar Duygulu, Automatic Image Captioning, Conference: Conference: Multimedia and Expo, 2004. ICME '04. 2004 IEEE International Conference ,Volume: 3.

[7] Oriol Vinyals, Alexander Toshev, Samy Bengio, Dumitru Erhan, Show and Tell: A Neural Image Caption Generator.