ISSN: 2366-1313

Image to Text: Detecting of Text based on Action using Convolutional Neural Network and Recurrent Neural Network

[1]Maguluri Reshma Venkat, [2]Thummala Sandhya, [3]Karicheti Sodhana Lakshmi, [4]Thanniru Sandhya, [5]Karavadi Geetha Lakshmi, [6]Vennela Alladi

[1]BTech Student, Dept.of CSE, QIS Institute of Technology, ONGOLE (A.P)

[2]BTech Student, Dept.of CSE, QIS Institute of Technology, ONGOLE (A.P)

[3]BTech Student, Dept.of CSE, QIS Institute of Technology, ONGOLE (A.P)

[4]BTech Student, Dept.of CSE, QIS Institute of Technology, ONGOLE (A.P)

[5]BTech Student, Dept.of CSE, QIS Institute of Technology, ONGOLE (A.P)

[6]Assistant Professor, Dept.of CSE, QIS Institute of Technology, ONGOLE (A.P)

Abstract: *Image Caption is a concept of gathering the right description of the given image on the internet use Computer Vision and natural language processing. The following is achieved using the Deep learning techniques called as convolution neural network and recurrent neural network. The dataset used for implementation is called as the Flickr8_k Dataset. The model uses the combination of convolution neural network (CNN) and the recurrent neural network (RNN) along with the Long short-term memory, which helps in extraction whereas the recurrent neural network helps in generation of the right text.*

Keywords: *Convolution Neural network (CNN), Recurrent Neural Network (RNN), Natural Language Processing, Computer Vision, ResNet, Long Short-Term Memory (LSTM).*

I. INTRODUCTION

Due to the rapid increase in the amount of data that is generated on the Internet, it has read to the problem of increase in the difficulty in organizing the data. Regardless of the availability of huge amount of raw data it is not feasible to convert them to resourceful data because of its highly unorganized nature. Image to text conversion play a vital role by automatically labeling large amount of data in a very short period of time which would otherwise be a tedious and time-consuming task involving a lot of labor. The various datasets available for the process are MS COCO, ImageNet, and Flickr8k are among others. The dataset chosen for this project is Flickr8k_Dataset due to its optimal size. The image captioning mainly happens in two phases. The first phase acts as an encoder in which the input image has been extracted of

ISSN: 2366-1313

relevant features[1]. This phase is implemented using the concept of Convolution Neural Networks which is one of the architectures mainly used in computer vision. The Pre-trained network called ResNet50 is a 34 layered architecture which has previously provided extremely reliable results in the field of image classification. The second phase comprises of Long Short-Term Memory. This is the decoder which basically converts the processed vector into an understandable and relevant sentence. The input image is encoded into an intermediate representation which comprises of information regarding the image which is subsequently decoded into sentences which best explains the image with minimum grammatical errors. One of the applications for image to text includes recommendation in editing applications. The image to text model helps in automating the process of providing captions for digital content and also accelerates this process. Another application is the assistance for visually impaired[2]. As we know there are many people in this world who haven't been naturally granted the boon to visually perceive and comprehend their surroundings. An image to text model can be widely used here to help such people to feel the surrounding in a much better way. In media and publishing houses there are vast amount of visual data which circulates in various forms. The image to text model helps provide and automatic subtitles for these data by automating as well as accelerating this process thus saves heaps of time and labor which would have otherwise gone in labeling this data. Last but not the least it can be used in social media posts. With artificial Intelligence rising to great heights, social media plays an important role and also increases the underline work for labelling and segregating these media files. An automatic image to text converter here would prove to be a major time saver[3].

High resolution of images has already shown the proof of exceptionally better performance at this, but the robust methodologies haven't yet been developed for earth imagery. Overcoming this the problem our aim is developing a combination of algorithm, encoder decoder architecture to caption these satellite images.

1. Review the info, which has detailed information about the labels and therefore the labeling process.

2. Downloading a sub-sample of the info to urge acquainted with how it's.Explore the sub-sample using python and exploratory data analytics.

ISSN: 2366-1313

3. Motivated by the burgeoning commercial and research interest in satellite images of Earth, we developed various models that are able to efficiently and accurately distinguish the content of such images.

Specifically, we trained deep convolutional neural networks (CNNs) to find out image features and used multiple classification frameworks including long short-term memory (LSTM) label captioning and binary cross entropy to predict multi-class, multi-label images.

II. LITERATURE SURVEY

Research in this paper focuses on the process of producing thematic from remote sensing of imagery for distinguishing images. Spectral bands non-analog integers are made to show spectral data. The data is made for non-analog distinguishing of pictures. In this paper, each pixel is distinguished through this spectral-data. Supervised and unsupervised are used for distinguishing images. This particular paper deals with the machine learning supervised distinguish mainly support vector machine, minimum distance, parallelepiped, and maximum likelihood

This paper deals with the strength of applying to NN computation to spatial image processing. The other AIM is to give a primary connecting of learning data in and normalize land area distinction outputs for conventional supervised and artificial neural net classes. ANN is trained to do land area classification of spatial clips of every dominant in the same way of supervised algorithms. This research is the base for creating applying weights for the future idea of software implications for ANN in the spatial image, earthly data preparation [4].

"Deep Visual-Semantic Alignments for Generating Image Descriptions"

We present a model that generates natural language descriptions of images and their regions. Our approach leverages datasets of images and their sentence descriptions to learn about the inter-model correspondences between language and visual data. Our alignment model is based on a novel combination of Convolutional Neural Networks over image regions, bidirectional Recurrent Neural Networks over sentences, and a structured objective that aligns the two modalities through a multimodal embedding. We then describe a Multimodal Recurrent Neural Network architecture that uses the inferred alignments to learn to generate novel descriptions of image regions. We demonstrate that our alignment model produces state of the art results in retrieval experiments on Flickr8K, Flickr30K and MSCOCO datasets. We then show that the

ISSN: 2366-1313

generated descriptions significantly outperform retrieval baselines on both full images and on a new dataset of region-level annotations.

"Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation"

Can a large convolutional neural network trained for whole-image classification on ImageNet be coaxed into detecting objects in PASCAL? We show that the answer is yes, and that the resulting system is simple, scalable, and boosts mean average precision, relative to the venerable deformable part model, by more than 40% (achieving a final mAP of 48% on VOC 2007). Our framework combines powerful computer vision techniques for generating bottom-up region proposals with recent advances in learning high-capacity convolutional neural networks. We call the resulting system R-CNN: Regions with CNN features. The same framework is also competitive with state-of-the-art semantic segmentation methods, demonstrating its flexibility. Beyond these results, we execute a battery of experiments that provide insight into what the network learns to represent, revealing a rich hierarchy of discriminative and often semantically meaningful features.

III. PROPOSED WORK

The model has two Neural Network Architecture convolution neural network and the other being Recurrent Neural Network. An exceptional type of RNN called as LSTM is used here which comprises of memory cell. The reason behind using LSTM is to keep the information for a long period time. We want our model to take an image as an input and give the output as the text description for it. Thus, the input image will be processed by CNN architecture which will give us the output for the CNN model and act as an input for the LSTM model. This will allow the LSTM to generate an output which is describing the image in text format. With this we have also use a pre-trained CNN called ResNet-50 architecture. This will help us to get the spatial information in the provided image

Model is built on Flickr 8k dataset which has optimal size, large enough to get a good accuracy.

• A combination of ResNet and LSTM is used.

• The image is extracted form a pre-trained model called ResNet in CNN which is acts as an input of the RNN for the generation of the caption

A. Dataset used for training the Model

Flickr8k_Dataset: Contains a total of 8092 images in JPG format with different

ISSN: 2366-1313

shapes and sizes. Of which 6000 are used for training, 1000 for test and 1000 for development. A new benchmark collection for sentence-basex image description for images that each image has five different captions which provide clear descriptions of silent entities and events.

B. Prepare the Text

All the sentences in the captions list is first extracted under the same list and further all the unique words of these sentences are extracted and added to a single list. The first step to be applied to convert the text into smaller units called tokens and then they are encoded with a numerical value. Here, the text is converted to a numerical form by assigning a number to all the unique words present in the text, and thus making it feasible to be used by the model. Sequencing in applied next on the tokenized text. Sequencing is where each sentence from the caption pool is taken and it is converted to a numerical list of the integer index of the respective word from tokenization. All the captions in the caption pool is of different lengths but the deep learning model requires an input of uniform shape to work on. To solve this problem, padding is used. Padding is where the shape of the input is considered to be of the length of the sentence with maximum words. The padding value used for our purpose is 0. Two delimiters "<start>" and "<end>" are used to mark the beginning and ending of the sentence.

C. The CNN-LSTM Architecture

We want our model to take an image as an input and give the output as the text description for it.Thus, the input image will be processed by CNN architecture which will give us the output for the CNN model and act as an input for the LSTM model. This will allow the LSTM to generate an output which is describing the image in text format. With this we have also use a pre-trained CNN called ResNet-50 (Fig. 2) [6] architecture. This will help us to get the spatial information in the provided image.

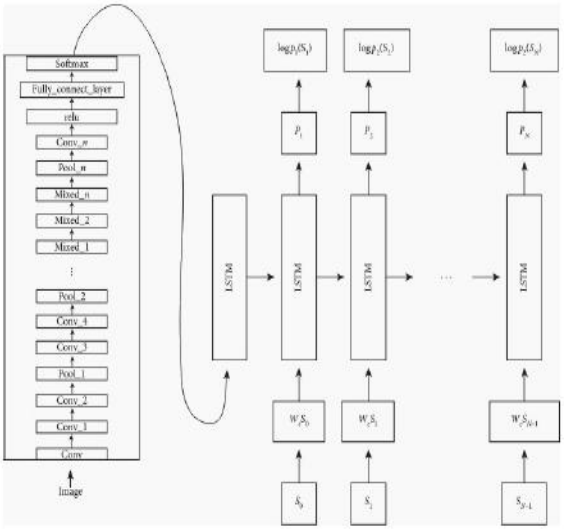
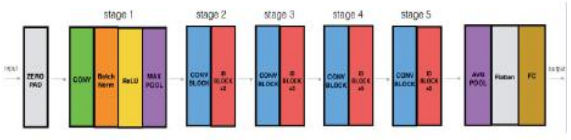


Fig.1 CNN-LSTM Architecture



ISSN: 2366-1313

Fig.2 Resnet-50 Architecture

D. ResNet 50

ResNet50 (Fig. 2.) is the architecture considered for this project. Residual Networks is the full form of ResNet which is one of the most popularly used deep models for transfer learning mainly for its capability to provide good results for really deep layers. The problem with earlier neural networks was that, when encountered with an extremely deep network, then a problem called vanishing gradients used to arise. Because of this problem, the accuracy would not improve after a certain point. This problem of vanishing gradient is taken care of in the ResNet architecture. ResNet50 was built on the ImageNet database which is a Collection of a vast number of images. This Architecture has 50 deep layers of convolutional neural networks. This network can recognize images over about 1000 categories with a very good accuracy. ResNet was also one of the first to bring out the concept of skip connection (Fig. 3.). Using this, the problem of gradient descent was addressed, by providing an alternative path which could be used by the gradient and hence minimizing the effect of vanishing gradient descent. It has a total of 50 layers in which 48 are convolutional layers. It has one max pooling layer and one average pooling layer

E. LSTM

The long-term dependency problem is dealt with in LSTM. They enable the model to have the capability of remembering the information for a longer period of time. The traditional recurrent neural network has a chain of modules which the LSTM has too, but in case of LSTM the neural network interacts in a special way. LSTM has played a vital role in the working of the image caption generation.

The model has two Neural Network Architecture (Fig. 1) convolution neural network (CNN) and the other being Recurrent Neural Network (RNN). An exceptional type of RNN called as LSTM is used here which comprises of memory cell. The reason behind using LSTM is to keep the information for a long period time

.

SYSTEM ARCHITECTURE

ISSN: 2366-1313

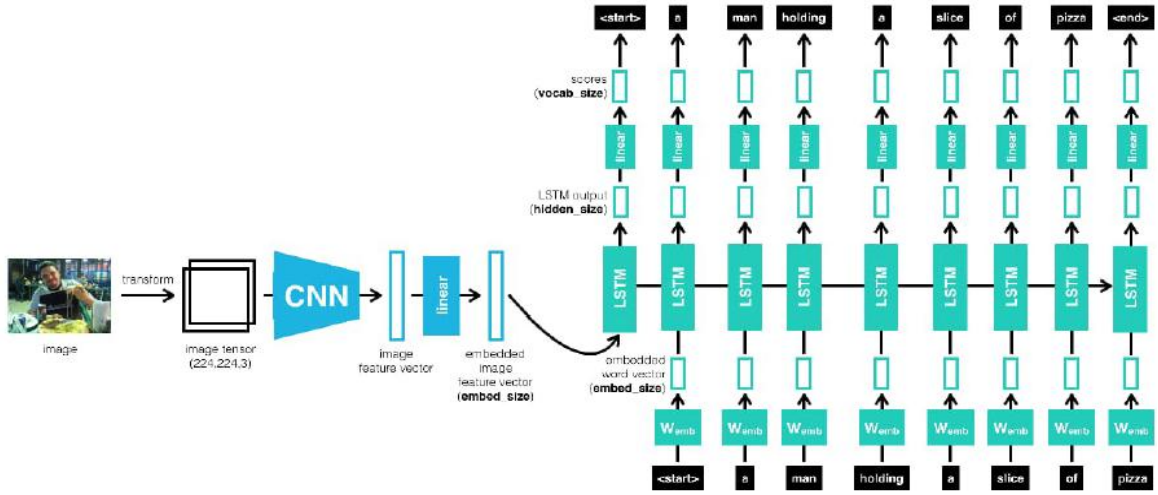


Fig. 3 System architecture

Data pre-processing

A dataset can be viewed as a collection of data objects, which are often also called as a records, points, vectors, patterns, events, cases, samples, observations, or entities. Data objects are described by a number of features that capture the basic characteristics of an object, such as the mass of a physical object or the time at which an event occurred, etc. Features are often called as variables, characteristics, fields, attributes, or dimensions. The data preprocessing in this forecast uses techniques like removal of noise in the data, the expulsion of missing information, modifying default values if relevant and grouping of attributes for prediction at various levels.

Deep Learning

Based on the split criterion, the cleansed data is split into 60% training and 40% test, then the dataset is Deep learning techniques called as convolution neural network and recurrent neural network. We can use this algorithm to find results.

IV. RESULTS

Here is the ideal result provided for image captions using CNN and LSTM. The selected, tested results were obtained using Flickr8k_Dataset. To get the result, we follow the special steps, which include loading libraries, loading registers, loading addresses as values and wizards as keys in a dictionary, mark-to-growth, review and validation. Download version Store directed subtitles. After performing these kinds of steps, the result of the title era can be seen in the output.

ISSN: 2366-1313

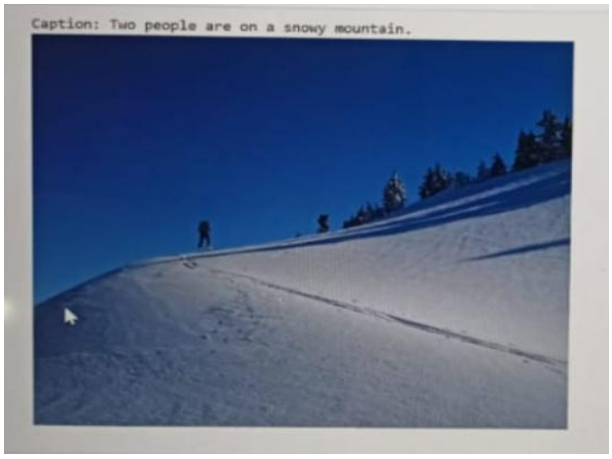


Fig.4 Two people are on a snowy mountain.

Fig.5 Baseball player in a red uniform is playing baseball.

V. CONCLUSION

Image recognition is a topic that draws a lot of attention to itself. It is one area that is very useful and thus attracts all advanced research work. This paper, Flickr8k_Dataset, was used as it proved to be the most appropriate length for investigation. There are many different datasets, such as MS COCO and Fickr30k, to name a few, that have also been shown to have significant and powerful consequences. The addition of CNN and LSTM is interesting because it provides an excellent result while also appearing in the form of a codec. CNN acts as an encoder that inserts the image and extracts features that may be needed for the image's feedback. When moving to LSTM, this converts ideas into sentence form reducing grammatical errors as much as possible. The framework used in this CNN challenge is the ResNet50 framework, a pre-qualified CNN module with 34 layers.

REFERENCES

[1] Thomas, Elsken; Jan Hendrik, Metzen; Frank, Hutter (2017-11-13). "Simple and Efficient Architecture Search for Convolutional Neural Networks"

[2] Elsken, Thomas; Metzen, Jan Hendrik; Hutter, Frank (2018-04-24). "Efficient Multi-objective Neural Architecture Search via Lamarckian Evolution".

[3] Springenberg, Jost Tobias; Dosovitskiy, Alexey; Brox, Thomas; Riedmiller, Martin (2014-12-21). "Striving for Simplicity: The All-Convolutional Net".

[4] Romanuke, Vadim (2017). "Appropriate number and allocation of ReLUs in convolutional neural networks". Research Bulletin of NTUU "Kyiv Polytechnic Institute". 1: 69–78.

ISSN: 2366-1313

[5] Bengio, Yoshua; Lamblin, Pascal; Popovici, Dan; Larochelle, Hugo (2007). "Greedy Layer-Wise Training of Deep Networks"(PDF). Advances in Neural Information Processing Systems: 153–160.

[6] "Using TPUs TensorFlow". TensorFlow. Retrieved 2018-11-14.

[7] S.Hochreiter, I.Schmidhuber," Long short-term memory," Neural computation, vol. 9, no. 8, pp. 1735-1780, 1997.

[8] J. Mao, W. Xu, Y. Yang, J. Wang, Z. Huang, and A. Yuille. Deep captioning with multimodal recurrent neural networks (m-RNN). arXiv preprint arXiv: 1412.6632, 2014.

[9] R. Kiros, R. Salakhutdinov, and R. Zemel. Multimodal neural language models. In International Conference on Machine Learning, pages 595–603, 2014.

[10] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell. Long-term recurrent convolutional networks for visual recognition and description. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 2625–2634, 2015.

[11] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 3156–3164, 2015