ISSN PRINT 2319 1775 Online 2320 7876
 Volume 12, Iss 1, 2023

Developing an Image Caption Generator with Big Data and Machine Learning Techniques

[1]Dr. Y. Krishna Bhargavi, [2]Dr Raghavender K V

Associate Professor, Department of CSE

GRIET, HYDERABAD, kittu.bhargavi@gmail.com

Assoc Prof CSE department

G NARAYANAMMA INSTITUTE OF TECHNOLOGY AND SCIENCE

drkvraghavender@gmail.com

Abstract: The Image Caption Generator is a sophisticated process that involves recognizing the contextual details of an image and annotating it with relevant captions, employing deep learning and computer vision technologies. Describing the content of an image automatically represents an intriguing and challenging task within the realm of artificial intelligence. This paper introduces an enhanced image captioning model that incorporates object detection, color analysis, and image captioning to autonomously generate textual descriptions for images. Image captioning holds significant utility across various applications, particularly in the analysis of extensive sets of unlabeled images. The proposed model, functioning as an encoder–decoder system for image captioning, utilizes VGG16 as an encoder and an LSTM (long short-term memory) network with attention as a decoder. Additionally, Mask R-CNN with OpenCV is employed for object detection and color analysis. The model is trained on the Flicker8k dataset, showcasing its versatility in applications such as aiding visually impaired individuals, improving image search capabilities, and facilitating human-computer interaction. The results substantiate the model's proficiency in comprehending images and generating coherent textual descriptions.

Keywords : *Convolution Neural Network (CNN), Recurrent Neural Network (RNN) , Long Short term Memory, Xception.*

1. INTRODUCTION

We are constantly exposed to images in our environment, on social media, and in the news. Photos can only be recognized by humans. Humans are able to identify images without the help of subtitles, but machines need to be taught how to recognize images first. Input vectors are used by the encoder-decoder architecture of image caption generator models to produce accurate and appropriate captions. This paradigm merges the fields of Computer vision, Deep learning and Natural language processing. Before describing anything in a natural language like English, one must first recognize and assess the context of the image. The CNN (Convolution Neural Network) and LSTM (Long Short Term Memory) are the two fundamental models on which our method is built. The derived application uses CNN as an encoder to extract features from the image and LSTM as a decoder to

ISSN PRINT 2319 1775 Online 2320 7876
Research paper Volume 12, Iss 1, 2023

organize the text and provide captions. Image captioning can be useful for a number of purposes, like supporting the blind with text-to-speech by providing real time information about the scene over a camera feed and enhancing social medical pleasure by rewriting captions for pictures in social feeds as well as spoken communications.A step in learning the language is helping kids identify molecules. Every picture on the internet should have a caption, as this would make exploring and indexing real photos faster and more precise. Biotechnology, business, the internet, and applications like self-driving automobiles, where it could describe the area around the car, and CCTV cameras, where the alarms might be sounded if any harmful activity is spotted, all use image captioning. This research article's major goal is to give readers a fundamental grasp of deep learning techniques.

Generating captions for images is a crucial job that relates to both the field of computer vision and the field of natural language processing. It is a very unbelievable development in artificial intelligence for a machine to mimic the human capacity for producing descriptions for visuals. The key issue in this work is to represent the relationships between things in the image in a language that is familiar to humans, such as English. Computer systems have historically generated text descriptions for photographs using predetermined templates. To generate lexically rich text descriptions, however, this approach does not offer enough variation. The improved effectiveness of neural networks has eliminated this flaw. Many cutting-edge models generate captions using neural networks by taking images as input and predicting next lexical unit in the output sentence.

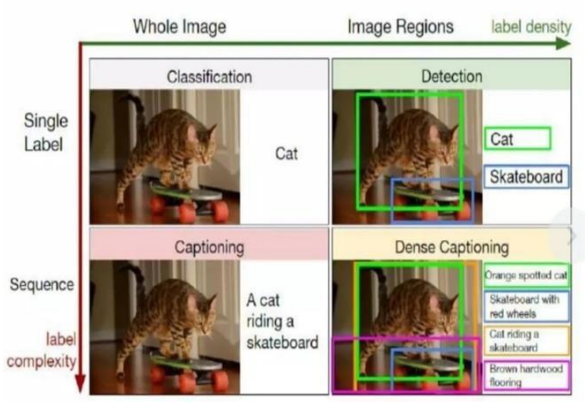


Fig 1 Example Figure

The project's goal is to produce a textual description from an image that is passed into it. The goal of the project is to use Deep Learning and Natural Language Processing to identify all the objects and attributes contained in an image, recognize the relationships between them, and then write captions summarizing each feature. The main objective is to build a image captioning generator so we will have a random image and our model will see the image and give us some captions.

2. LITERATURE REVIEW

Convolutional Image Captioning:

Image captioning is an important but challenging task, applicable to virtual assistants, editing tools, image indexing, and support of the disabled. Its challenges are due to the variability and ambiguity of possible image descriptions. In recent years significant progress has been made in image captioning, using Recurrent Neural Networks powered by long-short-term-memory (LSTM) units. Despite mitigating the vanishing gradient problem, and despite their compelling ability to memorize dependencies, LSTM units are complex and inherently sequential across time. To address this

ISSN PRINT 2319 1775 Online 2320 7876
 Volume 12, Iss 1, 2023

issue, recent work has shown benefits of convolutional networks for machine translation and conditional image generation. Inspired by their success, in this paper, we develop a convolutional image captioning technique. We demonstrate its efficacy on the challenging MSCOCO dataset and demonstrate performance on par with the baseline, while having a faster training time per number of parameters. We also perform a detailed analysis, providing compelling reasons in favor of convolutional language generation approaches.

Deep Compositional Captioning: Describing Novel Object Categories without Paired Training Data:

While recent deep neural network models have achieved promising results on the image captioning task, they rely largely on the availability of corpora with paired image and sentence captions to describe objects in context. In this work, we propose the Deep Compositional Captioner (DCC) to address the task of generating descriptions of novel objects which are not present in paired image-sentence datasets. Our method achieves this by leveraging large object recognition datasets and external text corpora and by transferring knowledge between semantically similar concepts. Current deep caption models can only describe objects contained in paired image-sentence corpora, despite the fact that they are pre-trained with large object recognition datasets, namely ImageNet. In contrast, our model can compose sentences that describe novel objects and their interactions with other objects. We demonstrate our model's ability to describe novel concepts by empirically evaluating its performance on MSCOCO and show qualitative results on ImageNet images of objects for which no paired image-caption data exist. Further, we extend our approach to generate descriptions of objects in video clips. Our results show that DCC has distinct advantages over existing image and video captioning approaches for generating descriptions of new objects in context.

Neural Machine Translation by Jointly Learning to Align and Translate:

Neural machine translation is a recently proposed approach to machine translation. Unlike the traditional statistical machine translation, the neural machine translation aims at building a single neural network that can be jointly tuned to maximize the translation performance. The models proposed recently for neural machine translation often belong to a family of encoder-decoders and consists of an encoder that encodes a source sentence into a fixed-length vector from which a decoder generates a translation. In this paper, we conjecture that the use of a fixed-length vector is a bottleneck in improving the performance of this basic encoder-decoder architecture, and propose to extend this by allowing a model to automatically (soft-)search for parts of a source sentence that are relevant to predicting a target word, without having to form these parts as a hard segment explicitly. With this new approach, we achieve a translation performance comparable to the existing state-of-the-art phrase-based system on the task of English-to-French translation. Furthermore, qualitative analysis reveals that the (soft-)alignments found by the model agree well with our intuition.

Show and Tell: Lessons learned from the 2015 MSCOCO Image Captioning Challenge:

Automatically describing the content of an image is a fundamental problem in artificial intelligence that connects computer vision and natural language processing. In this paper, we present a generative

ISSN PRINT 2319 1775 Online 2320 7876
Research paper Volume 12, Iss 1, 2023

model based on a deep recurrent architecture that combines recent advances in computer vision and machine translation and that can be used to generate natural sentences describing an image. The model is trained to maximize the likelihood of the target description sentence given the training image. Experiments on several datasets show the accuracy of the model and the fluency of the language it learns solely from image descriptions. Our model is often quite accurate, which we verify both qualitatively and quantitatively. Finally, given the recent surge of interest in this task, a competition was organized in 2015 using the newly released COCO dataset. We describe and analyze the various improvements we applied to our own baseline and show the resulting performance in the competition, which we won ex-aequo with a team from Microsoft Research, and provide an open source implementation in TensorFlow.

Show, Attend and Tell: Neural Image Caption Generation with Visual Attention:

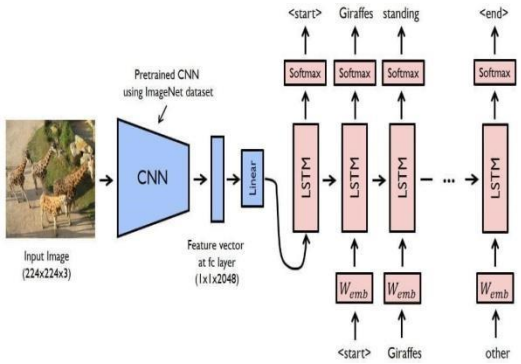
Inspired by recent work in machine translation and object detection, we introduce an attention based model that automatically learns to describe the content of images. We describe how we can train this model in a deterministic manner using standard backpropagation techniques and stochastically by maximizing a variational lower bound. We also show through visualization how the model is able to automatically learn to fix its gaze on salient objects while generating the corresponding words in the output sequence. We validate the use of attention with state-of-the-art performance on three benchmark datasets: Flickr8k, Flickr30k and MS COCO.

3. METHODOLOGY

For image captioning, we primarily employ two techniques: CNN and LSTM. As a result, we'll combine these architectures to create our image caption generator model. It's also known as the CNN-RNN model. The CNN algorithm is used to extract features from an image. We'll utilize the VGG16 model, which has already been trained. The input from CNN will be used by LSTM to help produce a description of the image. To define the structure of the model, we will be using the Keras Model from Functional API We used a small dataset consisting of 8091 images. For production-level models, we need to train on textual datasets larger than 40455 image captions which can produce better accuracy models.

Benefits:

- Recommendations in Editing Applications
- Assistance for visually impaired
- Social Media posts
- Self-Driving cars
- Robotics
- Easy to implement and connect to new data sources
- Reduce vehicle accidents
- Getting live captions from CCTV/Surveillance cameras
- Searching using image captions.



ISSN PRINT 2319 1775 Online 2320 7876
 Volume 12, Iss 1, 2023

Fig 2 System Architecture

Modules

To implement aforementioned project we have designed following modules

- Data exploration: using this module we will load data into system
- Processing: Using the module we will read data for processing
- Splitting data into train & test: using this module data will be divided into train & test
- Building the model - Resnet50, LSTM, Densenet121, Mobilenet, and Mobilenetv2
- User signup & login: Using this module will get registration and login
- User input: Using this module will give input for prediction.
- Prediction: final predicted displayed

4. IMPLEMENTATION

Algorithms:

CNN:

A Convolutional Neural Network (CNN) is a type of deep learning algorithm that is particularly well-suited for image recognition and processing tasks. It is made up of multiple layers, including convolutional layers, pooling layers, and fully connected layers.

The convolutional layers are the key component of a CNN, where filters are applied to the input image to extract features such as edges, textures, and shapes. The output of the convolutional layers is then passed through pooling layers, which are used to down-sample the feature maps, reducing the spatial dimensions while retaining the most important information. The output of the pooling layers is then passed through one or more fully connected layers, which are used to make a prediction or classify the image.

LSTM:

Long Short Term Memory is a kind of recurrent neural network. In RNN output from the last step is fed as input in the current step. LSTM was designed by Hochreiter & Schmidhuber. It tackled the problem of long-term dependencies of RNN in which the RNN cannot predict the word stored in the long-term memory but can give more accurate predictions from the recent information. As the gap length increases RNN does not give an efficient performance. LSTM can by default retain the information for a long period of time. It is used for processing, predicting, and classifying on the basis of time-series data.

Long Short-Term Memory (LSTM) is a type of Recurrent Neural Network (RNN) that is specifically designed to handle sequential data, such as time series, speech, and text. LSTM networks are capable of learning long-term dependencies in sequential data, which makes them well suited for tasks such as language translation, speech recognition, and time series forecasting.

VGG16:

ISSN PRINT 2319 1775 Online 2320 7876
Research paper Volume 12, Iss 1, 2023

VGG16 refers to the VGG model, also called VGGNet. It is a convolution neural network (CNN) model supporting 16 layers. K. Simonyan and A. Zisserman from Oxford University proposed this model and published it in a paper called Very Deep Convolutional Networks for Large-Scale Image Recognition.

5. EXPERIMENTAL RESULTS

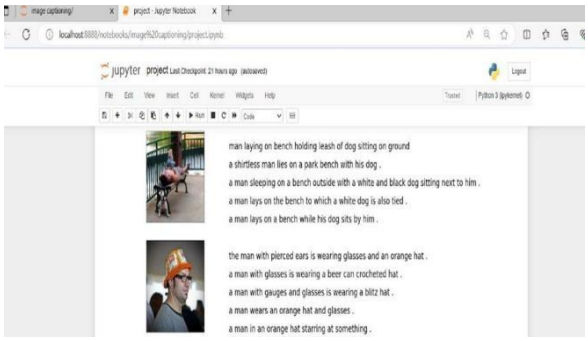


Fig 3 Output Screen

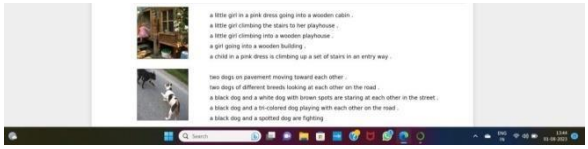


Fig 4 Output Screen

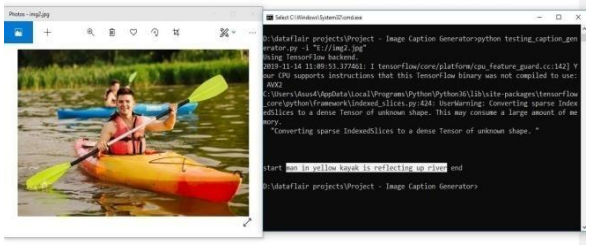


Fig 5 Output Screen

6. CONCLUSION

In this project we have learned and designed a technique of Image Caption Generator which will respond to User with captions or description based on an image. The Image Based Model extracts features of an image and the Language based model translates the features and objects extracted by image based model to a natural sentence. Image based model uses CNN whereas Language Based model used LSTM .The workflow is Data gathering followed by Pre-processing, Training model and Prediction. The ultimate purpose of an Image caption generator is to improve the social media platforms as well as in image indexing and for visually impaired persons with automated generated captions or description.

7. FUTURE SCOPE

1) Self driving cars-Automatic driving is one of the biggest challenges and if we can properly caption the scene around the car, it can give a boost to the self driving system.

2) Aid to the blind-We can create a product for the blind which will guide them travelling on the roads without the support of anyone else. We can do this by first converting the scene into text and then the text to voice. Both are now famous applications of Deep Learning.

3) Image captioning is used in a variety of sectors, including biology, business, the internet, and in applications such as self-driving cars wherein it could describe the scene around the car, and CCTV cameras where the alarms could be raised if any malicious activity is observed.

REFERENCES

[1] 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.

ISSN PRINT 2319 1775 Online 2320 7876
 Volume 12, Iss 1, 2023

[2] Lisa Anne Hendricks, Subhashini Venu gopalan, 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.

[3] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In International Conference on Learning Representations (ICLR).Shuang Bai and Shan An. 2018. A Survey on Automatic Image Caption Generation.Neuro computing.

[4] Vinyals O., Toshev A., Bengio S., Erhan D. "Show and tell: Lessons learned from the 2015 mscoco image captioning challenge." IEEE transactions on pattern analysis and machine intelligence. 2017 Apr 1;39(4):652-63.

[5] Xu K., Ba J., Kiros R., Cho K., Courville A., Salakhudinov R., Zemel R., Bengio Y. "Show, attend and tell:Neural image caption generation with visual attention." In International conference on machine learning 2015 Jun 1 (pp. 2048-2057).

[6] Liu C., Mao J., Sha F., Yuille A. L. "Attention Correctness in Neural Image Captioning." In AAAI 2017 Feb 4 (pp. 4176-4182).

[7] You Q., Jin H., Wang Z., Fang C., Luo J. "Image captioning with semantic attention." In Proceedings of the IEEE conference on computer vision and pattern recognition 2016 (pp. 4651-4659).

[8] Zhao S., Sharma P., Levinboim T., & Soricut R. "Informative Image Captioning with External Sources of Information," arXiv preprint arXiv:1906.08876, 2019.

[9] See A., Liu P. J., & Manning C. D. "Get to the point: Summarization with pointergenerator networks," arXiv preprint arXiv:1704.04368, 2017.

[10] Bahdanau D., Cho K., Bengio Y. "Neural machine translation by jointly learning to align and translate." arXiv preprint arXiv:1409.0473. 2014 Sep 1.

[11] Holger R. Maier and Graeme C. Dandy,Neural networks for the prediction and forecasting of water resource variables: a review of modelling issues and applica-tions,Environmental Modelling and Software,15,101{124,2000

[12] Avinash N. Bhute and B. B. Meshram,Text Based Approach For Indexing And Retrieval Of Image And Video: A Review,CoRR,abs/1404.1514,2014

[13] Keiron OShea and Ryan Nash,An Introduction to Convolutional Neural Net-works,CoRR,abs 1511.08458,2015

[14] Zachary Chase Lipton and David C. Kale and Charles Elkan and Randall C. Wetzel, Learning to Diagnose with LSTM Recurrent Neural Networks,4th International Confer-ence on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings , 2016

[15] Jurgen Schmidhuber,Deep learning in neural networks: An overview, Neural Networks,85{117,2015

ISSN PRINT 2319 1775 Online 2320 7876
 Volume 12, Iss 1, 2023

[16] Micah Hodosh and Peter Young and Julia Hockenmaier,Framing Image Description as a Ranking Task: Data, Models and Evaluation Metrics,J. Artif. Intell. Res,47,853{899,2013

[17] Tsung-Yi Lin and Michael Maire and Serge J. Belongie and Lubomir D. Bourdev and Ross B. Girshick and James Hays and Pietro Perona and Deva Ramanan and Piotr Dollar and C. Lawrence Zitnick,Microsoft COCO: Common Objects in Context,Computing Re-search Repository (CoRR),abs/1405.0312,2014

[18] Karen Simonyan and Andrew Zisserman,Very Deep Convolutional Networks for Large-Scale Image Recognition,Computer Science - Computer Vision and Pattern Recogni-tion,2014

[19] Polina Kuznetsova and Vicente Ordonez and Alexander C. Berg and Tamara L. Berg and Yejin Choi,Collective Generation of Natural Image Descriptions,359{368,The Association for Computer Linguistics,2012

[20] Siming Li and Girish Kulkarni and Tamara L. Berg and Alexander C. Berg and Yejin Choi,Composing Simple Image Descriptions using Web-scale N-grams,Proceedings of the Fifteenth Conference on Computational Natural Language Learning, CoNLL 2011, Port-land, Oregon, USA,220{228,ACL,2011