International Journal of Scientific Research in Science, Engineering and Technology
Print ISSN: 2395-1990 | Online ISSN : 2394-4099 (www.ijsrset.com)
doi : https://doi.org/10.32628/IJSRSET23102128

Scenerio-Based Image Generation Using Deep Learning

Dr. P. Shanthakumar[1] (M.E., Ph.D.), S. Kishor Kumar[2], S. Sharath[2]
Professor[1], B.Tech IT Scholar[2]
Department of Information Technology, Kings Engineering College, Sriperumbudur, Tamil Nādu, India

ARTICLEINFO

Article History:
Accepted: 10 April 2023
Published: 29 April 2023

Publication Issue
Volume 10, Issue 2
March-April-2023

Page Number
732-738

ABSTRACT

Regardless of the contents of the image, this model can determine its scenario. It works with open CV, Natural Language Processing, and websites. This system does a dictionary search on the image and displays the required results. Any type of image that the user possesses can be given the appropriate context. The computer's eyesight has made incredible strides in recent years. Using the latest technology, object detection is now simple and produces precise results. On the basis of the photograph, the goal is to create an exact situation. User-provided photographs may be utilised in a variety of online educational technology platforms to visualise and learn as well as for image searches. Resnet-50 algorithm is used and ResNet-50 is a convolution neural network that is 50 layers deep. You can load a pretrained version of the neural network trained on more than a million images from the ImageNet database we first encrypt the picture and the words that are kept in the scenario dictionary before the user searches using the image that he has. It will match the user-provided picture and the scenario dictionary after being encoded. After completing this procedure, it will discover the precise situation depicted in the image.

Keywords: Natural Language Processing, Object detection, User-provided photographs, Resnet-50algorithm, convolution neural network, situation depicted.

I. INTRODUCTION

Essentially, image searching is browsing where the words is replaced with graphics. In essence, we will have a picture and search utilising it to retrieve the results. The search query is expressed in the sample image.

Particularly for picture searches, there aren't many search phrases available. This basically removes the requirement for a user to choose from a list of keywords or phrases that could or might not provide the desired result. Users of image search may also learn about the popularity of a certain sample picture, related material, and other derived variations.

A scenario is nothing more than identifying the precise elements of a scene. We may use any example, such as a team playing roadside cricket or the stars in the night sky. Moving on to scenario matching, there are two key components to this process. Two are class and template, respectively. We have visual things in class such the human, apple, sun, and star. Templates will provide an illustration that combines a scenario and a set of classes.

A High computational cost is necessary to get more trustworthy findings and make the feature more strong and distinctive in terms of representation fusion of low-level visual characteristics.

In both supervised and unsupervised settings, a machine learning algorithm may be deployed by employing a training-testing framework.Deep neural network (DNN), which produce superior results at a high computational cost, are the focus of contemporary advances in image retrieval. The ability of the human eye to distinguish between diverse pictures based on colour makes colour one of the crucial low- level visual qualities. Images of the genuine item captured within the human visual spectrum can be recognised based on variations in colour.

II. LITERATURE SURVEY

The deep learning technique we used is called Generative Adversarial Network (GAN), which comprises of a generator and a discriminator. We also used Tensorflow, Numpy, NLTK, and Tensorlayer for text to image generation. In essence, Tensorflow is a library for machine learning. It compiles more quickly than other deep learning libraries. It also supports the use of GPU and CPU hardware. The NLTK (Natural Language Toolkit) tokenizer has been used to divide the text into smaller units, such as words, in this study. It makes it possible for the computer to review, prepare, and understand the textual content that the user submits.

Deep neural networks serve as generators and discriminators. While the Discriminator's objective is to identify accurate data, the Generator's objective is to deceive the Discriminator. Both Generator and Discriminator are in opposition to one another. The generator makes every effort to persuade the discriminator that the fake instances it generates are actual samples of data, and it also raises the likelihood of errors while the discriminator detects the true ones. As a result, these stages are repeated several times, greatly improving the training of both sub-models. To test if a discriminator can recognise actual data samples, those samples are used to train the discriminator. Again, bogus data is created and used to train the discriminator to determine if it can distinguish between real and fake photos.

Attribute learning is then applied to newly derived features from the CNN. This enables the classification of local region attributes using numerous trained Support Vector Regressors (SVRs), resulting in a collection of descriptive attribute names for each individual object area. For instance, insteadof using the default classification name, "apple," the algorithm can now refers to an area holding an apple as "a smooth green apple". The same idea has been used to expand the definition of "a person" to include "a young white male with brown hair," for instance. Overall, the system finds 26 and 25 characteristics forpersons and item descriptions, respectively.

To perform senior health and well-being tracking, the suggested system has also been combined with an intelligent user interface led by a 3D intelligent conversational assistant. By providing theintelligent agent with the system functions mentioned above, such as object and hazard detection, attribute classification, and background scene description generation, the system is able to learn about the immediate surroundings of users in order to warn them of hazards and generate alerts in the event of critical events like falling. In order to deliver more

individualised services and cope with the difficult open-ended natural human robot interaction, we also plan to integrate the suggested system with ahumanoid robot in the long run.

III. PROPOSED SYSTEM

Therefore, the suggested method seeks to address the aforementioned issues by producing in-depth descriptions of image contents using natural language. Through the use of people and object detectors, it will characterise the user's immediate surroundings, offer information about the user, alerts and queries, and provide generic descriptions. Our study focuses on a local region based strategy, which is more accurate than holistic strategies for scene classification and more explicitly relates to image regions of people and objects in a given image in order to retrieve more comprehensive information of the entire image. Our method deviates from analogous modern techniques in that we capture characteristics from theentire image rather than just a subset of it.

The proposed system consists of 4 main stages:

(1) Scene classification
(2) Object detection and classification (3)Attribute learning
(4) Sentence generation

IV. LEARNING ALGORITHM

The algorithm is StackGAN. The employed programming language is Python. The deep learning library used is called Torch. The website Better outcomes while browsing photographs was built using the Flask web framework. It displays several picture sources and references.We have used GAN CLS algorithm for training the discriminator and generator.

GAN-CLS Training Algorithm:
1. Input - minibatch images, matching text
2. Encode matching text description
3. Encode mismatching text description
4. Draw random noise sample
5. Generator will pass it to Discriminator
6. The pairs will be:
{Actual image, correct text}
{Actual image, incorrect text}
{Fake image, correct text}
7. Update discriminator
8. Update generator

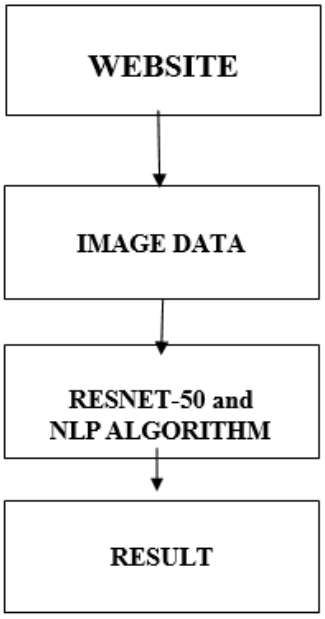

Figure 1: Learning Algorithm

NLP
Text is transformed using NLP into weights that are kept in dictionaries. Each time the user submits an image, the system compares it to the classes and a scenario contained in the dictionary and returns a response. That is the model's operational procedure. The whole model's necessary code was created in Python. The web framework in use here is Flask. The user and the model have a better interaction thanks to HTML. It will be helpful for those who are blind to understand the situation as well as for those who prefer to seek for images.

The three things the model we created can do It will compare the illustration and scenario and explain how

they relate to one another. It can accurately forecast the outcome of each image that a user posts. Additionally, it enhances the similar image search using the supplied image.

We'll create a vocabulary with categories and scenarios. There are objects in a class that are only pictures. The text encoder will convert the dictionary into a numerical weight value by utilising NLP. Using ResNet-50, the image encoder will transform the picture into weight.

The outcome will then be determined by the correlation between the two weights and the user's requirements.

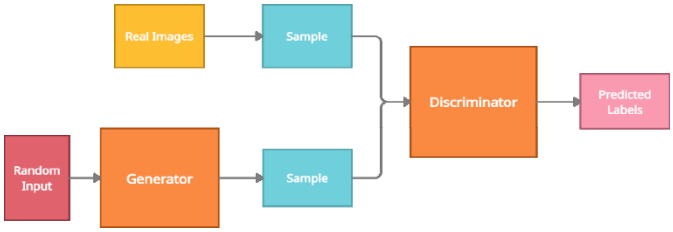

Figure 2: Pre-Processor

V. ARCHITECTURE

As was previously mentioned, the majority of prior techniques that sought to label or describe pictures depend on holistic approaches and deep learning techniques to gather characteristics and information from the complete image. This suggests that information might be lost because the architecture would not focus on specific regions or elements of the picture that might be necessary to fully describe the image. Regions enable the adding of more information to the picture descriptions by focusing the system's perspective and emphasising regions that current methods might not necessarily investigate. Our suggested study makes use of local regions classified using R-CNN under various conditions to get around this issue

Additionally, a lot of earlier state-of-the-art apps relied heavily on the fact that training and testing pictures were taken from the same area; otherwise, the quality of the results could be seriously harmed. However, in our study, cross-domain pictures or images outside of the scope can be evaluated upon by using the local R-CNN based object detectors.

The R-CNN serves as the foundation for the object detection used in this study. A conventional CNN is what the R-CNN is (with the addition of outputs that can forecast bounding box coordinates). This study uses the 200 object groups that can be detected, localized, and classified by the ImageNet ILSVRC13 dataset to train the R-CNN. Eight learned layers, including three completely linked layers and fiveconvolution layers, make up this CNN network structure.

The suggested system's initial step involves using R-CNN to find items in an image. It uses an SS algorithm to identify areas of images that may include persons, items, or both. The incoming picture is first scaled to a width of no more than 500 pixels. The SS algorithm blends the benefits of a thorough search with segmentation. It creates places using dimensions and visual characteristics. A greedy algorithm assesses similarity between areas and their neighbours and repeatedly groups similar regions. Repeating this procedure makes the entire picture into an area.

A collection of detections is created by concatenating the classification names. After being cropped, the areas and detections are once more labelled as either an item or a human. The sole purpose ofthis new label is to specify which group of attribute detectors (such as object or person attribute detectors)

FLOW DIAGRAM

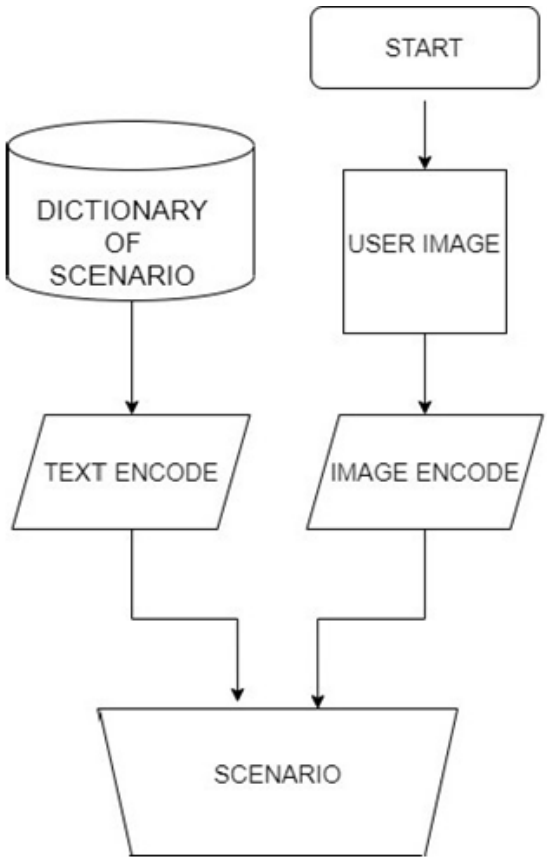

Figure 3 : image and text based scenario convertion

SCENE CLASSIFICATION

The hybrid Alex-Net deep CNN is used to perform scene categorization; it has been taught to recognise 1183 categories, including 205 scene labels and 978 object categories. However, because this mixed CNN network does not provide object location, it cannot be used in our work as an object detector. However, this network uses a GPU and runs on our GPU in about 0.2 seconds per picture. If speed is not a concern, this network can be used to create "cold brewed" Caffe, which uses CPU mode classification and generates results in about 2 seconds per picture.

A lot of computer vision apps used attributes, which can be thought of as high level features. Age, gender, and hair colour are examples of high level features in images. Size, colour, and texture are examples of high evel features in images that can be used to give image description creation systems strong descriptive capabilities.

Over the years, numerous methods for implementing sentence creation have been developed

Liuetal, use template matching to characterise video feeds. In their work, human actions are categorised and occurrences in the film are described using semantic level representations. To gather trait data, such as position, height, and motion, bounding boxes are used. This knowledge allowed for the determination of "subject-verb-object" and the creation of a narrative.

DESCRIBING VIDEOS

Small video segments are now included in the study for picture description creation. Yao et al.'s use of a 3D CNN that incorporates spatio-temporal data has enabled them to accomplish this.

In order to gather a unique feature representation from videos, this study uses a conventional CNN. To create sentences, this study incorporates LSTM into an RNN. Short segments' local motion descriptions are gathered by the 3D CNN and these characteristics are then merged with those from the standard 2D CNN.

Venugopalan et al.'s description of video segments also makes use of RNNs and large-scale deep CNNs. Initially learned on static pictures, the networks were refined using video description. This technique could have the drawback of producing a very basic single description that might not have enough information to adequately explain a video.

RESULT

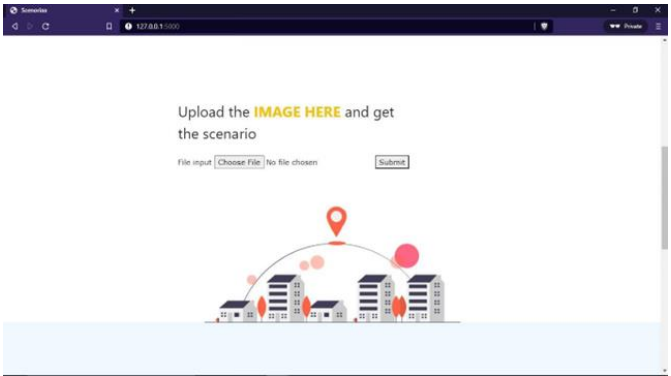

Figure 4: Sample output

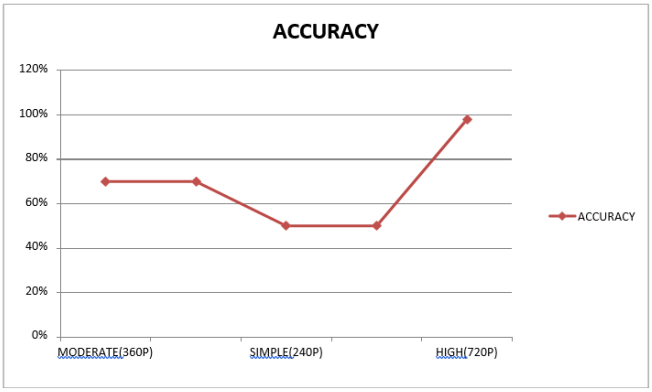

Figure 5: Accuracy Graph

VI. CONCLUSION

We looked at whether task-independent web-scale pre-training in NLP might be applied to other domains with success. We discover that using this formula leads to the emergence of comparable behaviours in the field of computer vision and talk about the societal ramifications of this line of inquiry. CLIP models acquire a wide range of skills during pretraining in order to maximise their training aim.

Then, by using natural language prompting, this task learning may be used to enable zero-shot transfer to several existing datasets. Although there is still considerable space for improvement, the performance of this technique may, at appropriate scale, compete with task-specific supervised models.

In this study, we suggested a local region-based method for creating multi-sentence picture descriptions. Regional features are included into template-based picture description production through the use of object recognition, attribute categorization, and relationship identification that are based on regions. Images beyond the scope can also be tested on using the local R-CNN based object detectors. The system contains fallen subject and hazard detection features as well, alerting users to hazards and dangers to help with geriatric care.

Experimental results show that our approach is capable of producing more precise descriptions of picture contents when compared to relevant state-of-the-art research. The performance of other related work is severely degraded when tested with such unrelated domain pictures, according to empirical studies, yet it exhibits tremendous flexibility and robustness when tested with cross-domain image datasets. Our approach performs better than other comparable systems for the development of out-of-scopedangerous and falling situation descriptions, especially in the area of geriatric care, whereas other related methods fail to produce any usable results.

VII. REFERENCES

[1]. A. Oliva, A. Torralba, "The role of context in object Recognition", Trends in Cognitive Sciences, 2007, 11(12): 520-527

[2]. N. Kumar, A. C. Berg, P. N. Belhumeur and S. K. Nayar, "Attribute and simile Classifiers for Face Verification," International Conference on Computer Vision (ICCV), pp. 365-372, 2009

[3]. O. Russakovsky and L. Fei-Fei, "Attribute Learning in Largescale Datasets." Trends and Topics in Computer Vision, pp. 1-14. 2010

[4]. R. Girshick, J. Donahue. T. Darrell, J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation." arxiv:1311.2524v5 2014

[5]. R. Girshick, "Fast R-CNN" arxiv:1504.08083 2015

[6]. S. Ren, K. He, R. Girshick, J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," arxiv:1506.01497 2015

[7]. L Zitnick, P. Dollar "Edge Boxes: Locating Object Proposals from Eedges" European Conference on Computer Vision, pp.391-405, 2014

[8]. N. Chavali, H. Argawal, A. Mahendru, D.Batra "Object-Proposal Evaluation Protocol is 'Gameable' arxiv:1505.05836 2015

[9]. P. Arbelaez, J. Pont-Tuset, J. Barron, F. Marques, J.Malik,"Multiscale Combinatorial Grouping", IEEE Conference on Computer Vision and Pattern Recognition, pp. 328-335, 2014

[10]. A. Humayun, F. Li, J.M Rehg, "RIGOR - Recycling Inference in Graph Cuts for generating Object Regions",IEEE Conference on Computer Vision and Pattern Recognition, pp. 336-343, 2014

[11]. O. Vinyals, A. Toshev, S. Bengio, D. Erhan, "Show and Tell: A neural Image Caption Generator" IEEE Conference on Computer Vision and Pattern Recognition, pp. 3156-3164, 2015

[12]. G. Kulkarni, V. Premaj, S. Dhar, S. Li, Y. Choi, A. C. Berg, T.L. Berg, "Baby Talk: Understanding and Generating Image Descriptions" IEEE Conference on Computer Vision and Pattern Recognition, 2011

[13]. K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. Zemel, Y. Bengio, "Show, Attend and Tell: Neural Image Caption Generation with Visual Attention" International Conference for Machine Learning (ICML), pp. 2048-2057, 2015

[14]. H. Fang, S. Gupta, F. Iandola, R. K. Srivastava, L. Deng, P.Dollar, J. Gao, X. He, M. Mitchell, J.C. Platt, C. L. Zitnick, G,Zweig, "From Captions to Visual Concepts and Back" IEEE Conference onComputer Vision and Pattern Recognition, pp. 1473-1482, 2015

[15]. A. Krizhevsky, I. Sutskever, G.E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks" Neural Information Processing Systems(NIPS), pp. 1097-1105, 2012

Cite this article as :

Dr. P. Shanthakuma1, Kishor Kumar, Sharath2, " Scenerio-Based Image Generation Using Deep Learning, International Journal of Scientific Research in Science, Engineering and Technology(IJSRSET), Print ISSN : 2395-1990, Online ISSN : 2394-4099, Volume 10, Issue 2, pp.732-738, March-April-2023. Available at doi : https://doi.org/10.32628/IJSRSET23102128
Journal URL : https://ijsrset.com/IJSRSET23102128