Image Classification and Annotation Using Deep Learning

Amrita Aman[1], Rajnish K Ranjan[2], Prashant Richhariya[3], Anita Soni[4]

[1]*Department of Computer Science & Engineering, Technocrats Institute of Technology (Advance), M.P., India.*
[2]*Department of Computer Science & Engineering, LNCTU, M.P., India*
[3,4]*Department of Computer Science & Engineering, Technocrats Institute of Technology (Advance), M.P., India.*

Abstract - *Deep learning has recently produced huge belief in the fields of AI. A lot of advanced research is running in this area image classification is one of them. In this work, we present a new deep learning model for image annotation and classification. Specifically, we propose a new probabilistic model that integrates image classification and annotation jointly. We also derive approximate inference and estimation algorithms based on various CNN and DNN methods, as well as efficient approximations for classifying and annotating new images. We demonstrate the performance of our model on the on CIFAR-10, CIFAR-100 and new dataset that we collected that consists of images and annotations of the same objects in different viewpoints. We show that our model is giving better result in compare to several baseline methods. In addition, we show that our model can be used for fast image annotation, which is an important task in computer vision. By comparing our model with a CNN-based method, we demonstrate that our model can be used to achieve comparable performance while using only a fraction of the time. Finally, we describe a novel, scalable implementation of our model. We insure that our implementation can be used to annotate all of the images in selected datasets and self-created benchmark dataset only in few seconds.*

Key Words: Classification, Annotation, Convolutional Neural Network, Deep Learning, Data set.

1.INTRODUCTION

The Computer Vision system focuses on identification and classification of the objects which are present in our surrounding as a key visual competence that will we use in the present scenario. One of the key difficulties in this area is image classification, which is going to be the most significant filed in future of Artificial Intelligence. Deep learning neural networks are among the most significant methods of machine learning which are used in this concept. Two independent problems that comes under computer vison are image classification and image annotation [3]. Our motivating intuition is that these two tasks should be connected. In this work, we develop a model based which is capturing input image and detecting semantic objects in the image. Based on the object's nature, model is used to annotate and name it. For detecting semantic objects it is important to extract various features of image like low level features color, shape and texture [2, 6] and mid-level features like motion and other salient features [14]. For new unknown images, our model provides predictive distributions of both class and annotation. Image classification is a computer vision technique for automatically categorizing images into a set of predefined classes. It aims to create a computer program that can automatically classify images into predefined categories. Images can be classified based on the objects [7] they contain or the scene they represent. Image classification has a wide range of applications. It can be used to categorize images of different objects in the real world, or to categorize images of different scenes. Image classification can be used for image retrieval, image detailing, image tagging, image annotation, etc. In the last few years, the demand for auto-annotation [13] methods has increased in the form of the recent outbreak of multimedia content and personal archiving on the Internet. Image classification is a multi-label classification problem that aims to associate a set of text with an image that describes its semantics. Basic working model of image classification can be seen in Fig-1 below.

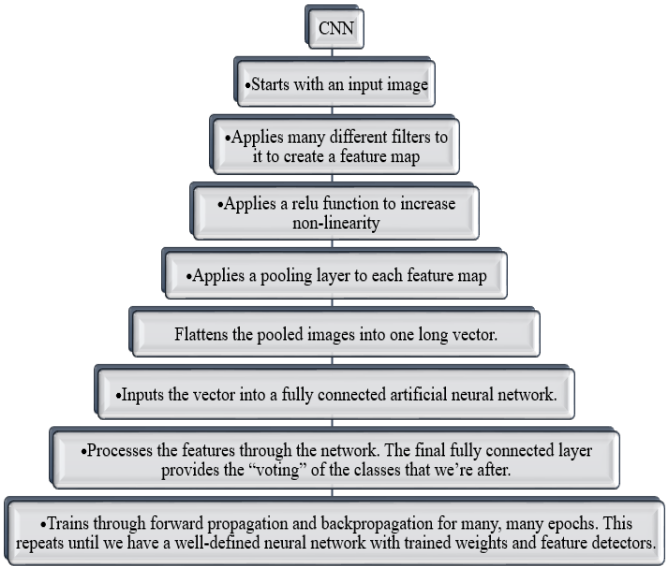

Fig -1: Working Flowchart of Convolutional Neural Network.

1.1 Motivation

Image classification is major challenge today. In our day-to-day life scenario, an object can be categorized into multiple classes. That can be a newspaper column which is tagged as "political", "election", "democracy"; an image can contain "tiger", "grass", "river"; and many more. These are some samples of multi-label classification, which accommodate with the task of linking multiple labels with single form of

data. This is a complicated problem because it needs to consider the complex correlations which is shared by different labels.

2. LITERATURE REVIEW

On the face of it, image classification [1] seems simple. You show an image and the computer tells you what it is. But in actuality, classification is very difficult. The computer needs to be able to tell the difference between a picture of a person and a picture of a dog. It needs to be able to tell the difference between an image of a human face and an image of a human body. It needs to be able to tell the difference between an image of a cat and an image of a person. As the number of classes increases, the difficulty of classification also increases. The more classes there are to classify, the more difficult the classification becomes. The more classes there are to classify, the more difficult the classification becomes. So, how do we do this? We used one of the most useful tools in computer science: machine learning. Machine learning is a field of computer science that involves developing algorithms that can learn and improve on their own. Machine learning is used in a variety of fields, including web crawling, natural language processing, speech recognition, and image classification [4]. When we talk about image classification, we are essentially talking about training a machine learning algorithm. This is a process of training a machine-learning algorithm to recognize a specific class. which comes under deep learning. Deep learning [5] is a subset of machine learning that is concerned with the representation and use of data. Deep learning is usually implemented using neural networks. A neural network is a mathematical model that simulates the human brain. It is comprised of three layers: an input layer, a hidden layer, and an output layer. Each layer consists of a set of neurons that can be connected to one another. Each neuron consists of a weight and a threshold. The weight represents the strength of the connection [11] between the neuron and the neuron next to it. The threshold represents the average strength of the connection of the neuron to the neurons on either side. Architecture of neural network can be understand like- Data is received by the input layers and transmitted to hidden layers. Number of hidden layers depends on the model. As increasing the number of hidden layers, increases the accuracy but computation cost also increases. The hidden layer performs a process called non-linear transformation. It maps the data into a new space, where the data is more suitable for training the neural network. The hidden layer then passes the data to the output layer. The output layer performs a process called linear transformation. It maps the data into a space where the data can be used to make predictions.

The first step in training a machine learning algorithm is to define the problem. In this case, we are trying to train a machine learning algorithm to recognize images of human faces. So our task is to define a problem that we can solve. The next step is to define the features of the problem. The features are the characteristics that define the class of objects we are trying to identify. For example, the features of a face may include: the eyes, nose, lips, and hair. Once we have defined the problem and features, we can start training the algorithm. We start by using a set of features that can be extracted from an image. These features are used to describe the image in some way. For example, the computer may use the color of the image, the brightness, the contrast, the texture, or the shape of the image [6, 14]. Features are represented by numerical values, which are then fed into a classification algorithm. We have a set of feature vectors for each image. The classification algorithm then uses these feature vectors to make a prediction about the image. Usually, the predicted label is the most common one in the training data. The training data is a collection of feature vectors and the corresponding predicted labels. The training set is used to learn the parameters of the classifier so that it can predict the labels of new examples, There are many different classification algorithms, each with their own strengths and weaknesses. One of the most popular classification algorithms is the CNN. Convolutional Neural Network [8, 9] has many layers in its implementation. Pooling layer (Fig-2) is one of them which is used to reduce the dimension. It helps to reduce the number of parameter of data and hence require less computation. Size of the kernel in pooling layer can vary as per the requirements. In Fig -2 it is of 2 X 2 kernel which is 4 X 4 data matrix into 2 X 2. Concept behind reduction of parameter is picking the highest magnitude and ignore other smaller magnitude. Then kernel window shift forward for picking another highest magnitude value in that particular window. And this way number of resulting parameter reduced.

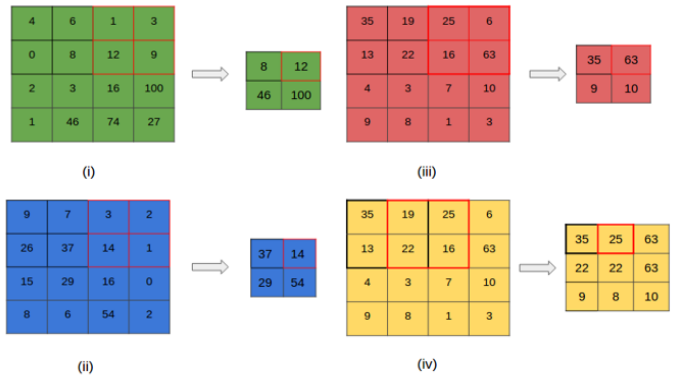

Fig -2: Pooling Process for different dimensions.

Convolutional neural networks are a type of neural network that have become popular in recent years. They have been applied to a variety of problems, including: speech recognition, image classification, and object detection. The CNN is a deep learning [12] algorithm that can classify images and videos. It has been trained to detect objects in images. It does this using a convolutional neural network.

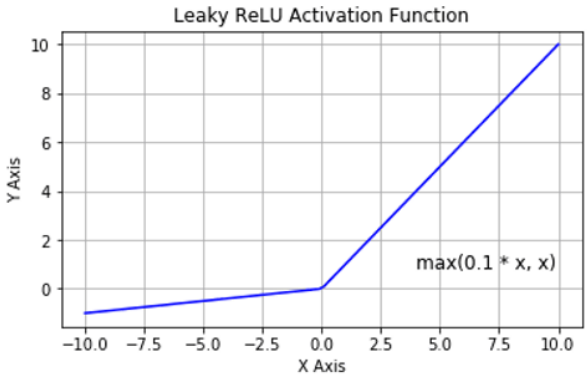

Fig -3: Leaky ReLU Activation Function Graphical Representation.

3. PROPOSED METHODOLOGY

We use convolutional filters to extract features from the image. The convolutional filters [10] are applied to the input image to generate a feature map, which is a matrix of numbers that represents the features in the image. "Block Diagram for proposed techniques are shown in below Fig - 4." The number of columns in the feature map is equal to the number of filters that were used. The number of rows in the feature map is equal to the number of pixels in the input image. Each element in the feature map is a number that represents how much the filter affects the input image. The filter is applied to the input image and the feature map is generated. The output of the filter is then passed through a nonlinear activation function, such as a rectified linear unit (ReLU). Many other linear and non-linear activation function exists and each has its own advantages and disadvantages. One of the most important feature of ReLU is that, it does not activate all neurons in same time. But the problem with ReLU is that- It is unbounded & Not differentiable at zero. Leaky ReLU (Fig-3) can be used to overcome ReLU problems. The output of the activation function is then multiplied by its corresponding weights, and the sum is passed through the activation function again. The weights are stored in a matrix that is called a weight matrix. The weights are the numbers that are used to multiply the input image and generate the feature map. The weights are also used to generate the feature map. If the weights are large, then the corresponding elements in the feature map will also be large. If the weights are small, then the corresponding elements in the feature map will also be small. In general, the output of the activation function is used as the input to the next layer, which in turn generates a feature map. This process is repeated until the feature map is generated. The final feature map is then used to classify the objects in the image. If we use a convolutional neural network to classify objects, it is called a two-stage detector. The first stage is the object detection stage.

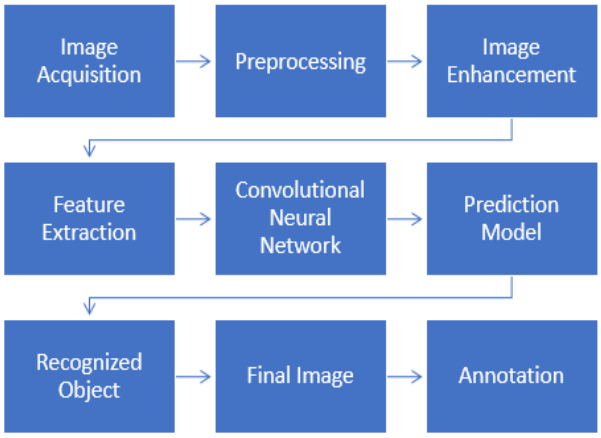

Fig -4: Basic Block Diagram for Classification and Annotation.

This stage is a convolutional neural network that is used to identify objects in an image. The second stage is the classification stage. This stage is a fully connected neural network that is used to classify the objects that were identified in the first stage. The first stage is also called the proposal stage. It is the stage that generates the proposed regions. The second stage is also called the classification stage. It is the stage that classifies the proposed regions. Future of image classification is coming with coming year of advanced technologies; computer vision techniques have improved very rapidly and novel deep learning-based detection methods have been proposed. With coming year of advanced technologies, computer vision techniques have improved very rapidly and novel deep learning-based detection methods have been proposed.

4. RESULT AND DISCUSSION

4.1 Dataset

Proposed model for classification and annotation has been tested for several datasets like CIFAR-10, CIFAR-100, COCO and self-generated dataset. CIFAR-10 is the set of 10 different classes like- airplanes, birds, cats, dogs, etc.; with more than 60000 color images whereas, CIFAR-100 has 100 different classes with more than 600 images in each class. COCO (Common Object in Context) dataset is used for classification, recognition, segmentation, object detection and for many other purposes. It has more than 330K images with 200K+ annotated images.

4.2 Performance Metric

Performance of the model has been calculated based on classification performance metric shown in Table -1. Where TP, FP, FN and TN abbreviate True Positive, False Positive, False Negative and True Negative respectively. Actual value and predicted value represented as rows and column respectively.

Table -1: Classification Performance Metric

1: True; 0: False; TP: True +Ve; FP: False +Ve FN: False –Ve; TN: True –Ve.		Actual	
		1	0
Predicted	1	TP	FP
	0	FN	TN

Accuracy of the module is calculated by the following formula-

Accuracy = (TP + TN) / (TP + FP + FN + TN)

4.3 Classification and Annotation

Convolutional Neural Network is used for classification and one can explore its working process through Fig -5. Self-generated input data has been passed for various CNN layers like convolutional layer, pooling layer and fully connected layer. Each layer has its own features to extract some knowledge for input image. This work is identifying all semantic objects in background/ foreground in image and

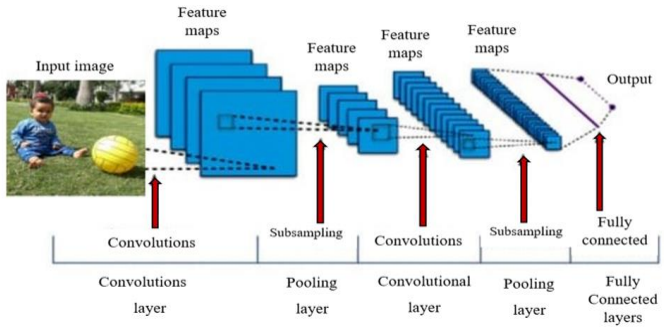

Fig -5: CNN Layers for input image.

based on identified objects, model is recognizing it with various existing features to annotate and name it. Fig -6 (Self-generated data) and Fig -7 (CIFAR-10 Dataset) tells everything about output of this module. In Fig -6, A sports ball and a baby (person) can be manually visualize in input image and model is classifying it accurately and annotating perfectly. In Fig -7, A train and a pole can be manually visualized but this module is identifying semantic object as train and annotating it.

Fig -6: Classification of Objects and Annotating it for Self-Generated Data.

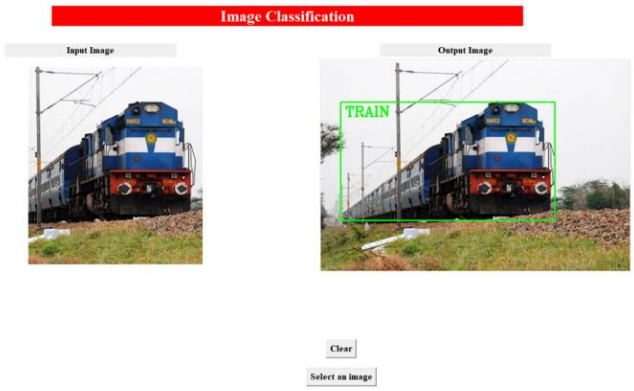

Fig -7: Classification of Objects and Annotating it for CIFAR-10 Dataset.

4.4 Work Comparison

Develop model has been tested on various existing standard datasets as well as on self-generated benchmark datasets. Accuracy of the model for some of standard datasets like CIFAR-10, CIFAR-100 and self -created dataset are tabulated in Table -2. Its performance is 98.78% for benchmark self-created data and 96.78% for CIFAR-10 & 84.96% for CIFAR-100 datasets are recorded. Performance of some standard model in the field of classification like DenseNet and Drop-Activation has been tabulated for the same dataset. Both model almost performing same over CIFAR-10 but in case of CIFAR-100 Drop-Activation model is little bit better classifier in compare to DenseNet model.

Table -2: Result Obtained from Different Model

Model	Dataset	Accuracy (%)
DenseNet	CIFAR-10	96.54
	CIFAR-100	82.82
Drop-Activation	CIFAR-10	96.55
	CIFAR-100	83.80
Proposed Approach	CIFAR-10	96.78
	CIFAR-100	84.96
	Self-Created	98.78

5. CONCLUSIONS

This work has given an insight into image annotation and classification using deep neural networks. As deep learning applications to image classification are integral part of machine learning that shows the power of artificial intelligence. Such deep learning techniques are used to solve complex problems as human brain does automatically. Here we have used CNN techniques along with Leaky ReLU activation function to train the model. Training has been done using forward propagation as well as backward propagation for many epochs. Model has been trained and tested over datasets like COCO, CIFAR-10, CIFAR-100 and other benchmark datasets. We used precision and recall metric to find the accuracy of the model. Results for classification and annotations can also be visually seen and

assess the accuracy of our model. Proposed work has been compared with various latest work done in this area, few of them are like- DenseNet, Drop-Activation, etc. Finally, we assure about our model that, its performance is with 98.78% accuracy for benchmark self-created data, 96.78% for CIFAR-10 and 84.96% for CIFAR-100 datasets. We are carrying forward this work with high hope and huge expectation for optimal accuracy using combined deep learning techniques CNN and RNN.

REFERENCES

[1] Rawat, Waseem, and Zenghui Wang. "Deep convolutional neural networks for image classification: A comprehensive review." Neural computation 29.9 (2017): 2352-2449.

[2] Kumar, Gaurav, and Pradeep Kumar Bhatia. "A detailed review of feature extraction in image processing systems." 2014 Fourth international conference on advanced computing & communication technologies. IEEE, 2014.

[3] Ojha, Utkarsh, Utsav Adhikari, and Dushyant Kumar Singh. "Image annotation using deep learning: A review." 2017 International Conference on Intelligent Computing and Control (I2C2). IEEE, 2017.

[4] Pin Wang , En Fan , Peng Wang , Comparative Analysis of Image Classification Algorithms Based on Traditional Machine Learning and Deep Learning, Pattern Recognition Letters (2020).

[5] Obaid, Kavi B., Subhi Zeebaree, and Omar M. Ahmed. "Deep learning models based on image classification: a review." International Journal of Science and Business 4.11 (2020): 75-81.

[6] Rajnish K. Ranjan and Anupam Agrawal, "Video Summary Based on F-Sift, Tamura Textural and Middle level Semantic Feature," Procedia Computer Science, 12th International Conference on Image and Signal Processing 2016, Volume 89, pp. 870-876.

[7] Sudharshan, Duth P., and Swathi Raj. "Object recognition in images using convolutional neural network." 2018 2nd International Conference on Inventive Systems and Control (ICISC). IEEE, 2018.

[8] Abu, Mohd Azlan, et al. "A study on Image Classification based on Deep Learning and Tensorflow." Int. J. Eng. Res. Technol 12.4 (2019): 563-569.

[9] Shorten, Connor, and Taghi M. Khoshgoftaar. "A survey on image data augmentation for deep learning." Journal of big data 6.1 (2019): 1-48.

[10] Sun, Yanan, et al. "Evolving deep convolutional neural networks for image classification." IEEE Transactions on Evolutionary Computation 24.2 (2019): 394-407.

[11] Bhawsar Y., Ranjan R.K. (2021) Link Prediction Computational Models: A Comparative Study. In: Tomar R.S. et al. (eds) Communication, Networks and Computing. CNC 2020. Communications in Computer and Information Science, vol 1502. Springer, Singapore.

[12] Aa, Vasuki, and Govindaraju Sb. "Deep neural networks for image classification." Deep Learning for Image Processing Applications 31 (2017): 27.

[13] Murthy, Venkatesh N., Subhransu Maji, and R. Manmatha. "Automatic image annotation using deep learning representations." Proceedings of the 5th ACM on International Conference on Multimedia Retrieval. 2015.

[14] Ranjan R.K., Bhawsar Y., Aman A. (2021) Video Summary Based on Visual and Mid-level Semantic Features. In: Tomar R.S. et al. (eds) Communication, Networks and Computing. CNC 2020. Communications in Computer and Information Science, vol 1502. Springer, Singapore.

BIOGRAPHIES

Amrita Aman is a M.Tech candidate in the Department of Computer Science and Engineering at TITA Bhopal. She graduated in CSE (HIT-K) in 2013 and served as an Assistant Manager in Bank of Baroda for 4 years (2015-2019).

Rajnish K Ranjan has master's degree in Computer Sc. & Engineering from IIIT Allahabad (2016). He is currently Lecturer in GWPC Bhopal in Department of Computer Science & Engineering and Pursuing Ph.D. from LNCTU, Bhopal. His research area is Machine Learning (Deep Learning), Internet of Things (IoT) and Computer Vision.

Prashant Richhariya worked as an Assistant Professor in TITA Bhopal. He is having an experience of 18 years in Teaching. His research area is computer vision, image processing, Data mining. He is having total 5 patient, 18 research papers.

Dr. Anita Soni is Professor in Department of CS (TIT-Advance). She has more than 15 years of teaching and research experience. She has 3 Patient, 3 Book Chapters and 15 Research Paper.