TRANFORMERS BASED MULTI-LABEL IMAGE CLASSIFICATION AND NAMING

Jerin Jacob[1], Dr. G Kiruthiga[2]

[1] *Student, Dept of Computer Science and Engineering, IES College of Engineering, Kerala, India*
[2] *Associate Professor, Dept of Computer Science Engineering, IES College of Engineering, Kerala, India*

ABSTRACT

In recent years, the use of Machine Learning (ML) which is a type of Artificial Intelligence (AI) has increased rapidly. ML allows software applications to become more accurate in predicting the output without explicitly programmed to do so. Classification is a type of ML which categorizes given data into different classes and is used in speech recognition, face detection etc. Here image classification is considered. Multi-label classification is method of predicting set of labels according to attributes, objects present in the given image. Transformers are classification models which was using in Natural Language Processing (NLP). Now it is using in image classification. The transformer is given the data after feature extraction. Then it will be capable of predicting the labels in the image. Here in this paper, we are considering Vision Transformer (ViT). It is a model for multi-label image classification which can exploit the complex dependencies between visual features and labels. Here the concept of Multilayer Perceptron (MLP) is used to extract the features from the images. In this approach the vision-transformer is trained to predict the masked labels from the input given to it. Here the dataset used for training is the CIFAR-10. The core of this method is a label mask training objective. This model can represent label state explicitly during training. Thus, this model can be effectively applicable in medical image recognitions, wild animal recognitions and so on. This model can work at an efficiency of about 84% in its prediction accuracy.

Keyword: - *Multilabel images, Vision Transformer (ViT), Multilayer Perceptron (MLP)*

1. INTRODUCTION

In machine learning, analysis of data is the main task. Neural Networks emerged as a most effective tool for the analysis of the big data and for the feature extraction of the same. In this type of networks, there will be several layers for the purpose of feature extraction. And these extracted feature details will keep in a file and use this file for the prediction of next output. This is the basic working of ML. The following are main portions of our proposed model.

1.1 Multilayer Perceptron

There are many types of neural networks available which are proposed by researchers and tweaks of the existing networks. In this model a classical type of artificial neural network called Multilayer Perceptron (MLP) [5] is used. It has several layers of neurons and each are connected using directed graph, so signal path in neurons is one way. MLP uses a supervised learning technique namely backpropagation. The input is fed to one layer, it is processed through the hidden layers which may be one or more and it is used for levels of abstractions within it. Finally, predictions are made at the output layer which is also called as the visible layer. It is very flexible and can be used to learn the mapping from input to output. In this proposed model MLP along with the transformer is used for the feature extraction from the dataset.

1.2 Vision Transformer

Vision Transformer (ViT) [6] has emerged as a competitive to the Convolutional Neural Network (CNN) in the last years. Currently CNN has the monopoly in the computer vision and ViT outperformed CNN by 4 times in its computational efficiency and accuracy. Natural Language Processing (NLP) was the area where Transformer models used. Now it is using in computer vision too. A transformer is applied to a sequence of image patches for image classification task.

1.3 Why ViT

Images in real life can have multiple objects in one image itself. Multilabel image classification is basically a visual recognition and predict the set of labels in that image. Labels include objects, actions or attributes in the given image. When using the Convolutional Neural Network (CNN), model for the feature extraction, after several number of encodings, several features may loss. But in the case of transformer there will be a self-attention to the features after any number of encoding and no data will be lost after that. Thus, we chose the ViT network for doing this prediction.

2. RELATED WORK

Classification transformer (C-Tran): Classification Transformer [1] which is proposed as a multi-label classification which can work in regular inference, inference with partial labels and inference with extra labels. This formulation allows to easily input the image features with labels to the transformer encoder. Transformer encoders are of order invariant and it allow any type of dependencies between all features and labels learned. For image comparison they have used the ResNet-101 model and there is a total of 324 feature embedding vectors. C-Trans use four attention headers for the transformer encoding.

Graph Convolutional Network: This model builds a directed graph over the object label and Graph Convolutional Network (GCN) [3] is learned to map this label to set of inter dependent object classifiers. The Overall performance was good and it was on the classifiers of ResNet on MS-COCO.

Multilabel image classification: Multilabel image classification [4] is gaining popularity due to its relevance in world wide applications. There were some errors in ImageNet on the prediction. It was not due to the problem of feature extraction but ImageNet was annotated to single label while some image depicts more than one object or label.

3. PROPOSED SYSTEM

In the proposed system, the feature extraction is done by the MLP model with the Vision Transformer (ViT), on the dataset of CIFAR-10 which is commonly available for research purpose. The image which is fed as the input will be patched into 16x16 pixel [2] images and fed to the ViT. Since the image has been patched into small ones, there is a chance of getting unordered while considering for the prediction by the encoder. If the order is changed, the image will be get changed and it may become non understandable too and it may not have any match with the image given as the input. So each patches are bound together with a positional encoding and pass those to the transformer encoder. Unlike the CNN, transformers are based on the self-attention layers which means no feature will be lost after repeated feature extraction process. CNN may lose several features after several extraction.

The data after the transformer encoder is fed to the MLP model. MLP is used to vectorize the image patches to process by the vision-transformer. And from there the vision transformer can predict the labels which are in the image.

3.1 Tools Used

COLAB: COLAB is a product developed by google research. It represents ‘colaboratory’ and it allows anyone to write and execute arbitrary python through browser. Machine learning is a heavy task which need more computational power, especially a powerful GPU (Graphic Processing Unit) for the process of feature extraction and thus it is not suitable to run this work in normal PC. It is not impossible but it need comparatively more time to execute these files. Here COLAB provide a high-performance GPU to its users for this type of processing. And we chose COLAB as our platform.

CIFAR-10: It is a dataset available for the research purpose for machine learning experiments. Dataset is a complicated thing which include thousands or ten thousand of images of different classes and which is available for the learning purposes.

Tensorflow, Keras: Tensorflow is an open-source platform and it is a library for different machine learning tasks. Keras is a high-level neural network library working on the top of Tensorflow. Keras is highly-productive interface in solving ML problems. Keras is comparatively simple because it is developed on Python language. Because of this high-productivity and ease of use we chose the Keras.

3.2 Implementation

Here all the operations are done as layers. When we give an image to the model, the first layer patches that image into 16x16 pixels and this patching is done by a layer. The images which are patched are given to the transformer encoder after embedding with the position information. This information is very important since if the position of patches have changed, then the prediction may go wrong exactly.

Figure 1: Converting image to patches of 16x16 pixel size [2]

The transformer encoding is done at several internal hidden layers of the transformer and after that the patch is given to the MLP and after the extraction by MLP is completed, the labels will be matched with the image.

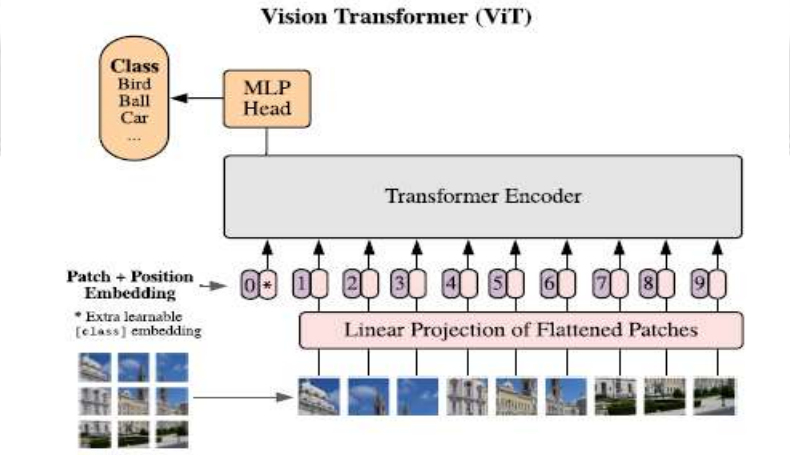

Figure 2: Architecture of ViT [2]

3.3 Advantages of the proposed system

1. Since the model used here is transformer, comparatively faster
2. quality of result and accuracy is more a than other models

4. CONCLUSIONS

We propose, Transformers based multilabel image classification and naming is a novel and flexible model. This approach is easy to implement. This model can learn simple adaptive interactions through attention and discovers how labels attend different parts of the image. This model performs well than the normal classification techniques using today. By this model we have provided a qualitative and quantitative analysis. This model can be extended and can be used in medical field for the more accurate image predictions.

5. REFERENCES

[1]. "General Multi-label Image Classification with Transformers"
Jack Lanchantin, Tianlu Wang, Vicente Ordonez, Yanjun Qi University of Virginia

[2]. "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale."
Alexander Kolesnikov Alexey Dosovitskiy Dirk Weissenborn Georg Heigold Jakob Uszkoreit Lucas Beyer Matthias Minderer Mostafa Dehghani Neil Houlsby Sylvain Gelly Thomas Unterthiner Xiaohua Zhai. ICLR (2021)

[3]. Zhao-Min Chen1, Xiu-Shen Wei Peng and Wang Yanwen Guo "Multi-Label Image Recognition with Graph Convolutional Networks" - 978-1-5090-0620-5/16/$31.00 c 2016 IEEE

[4]. Pierre Stock and Moustapha Cisse. "Convnets and imagenet beyond accuracy: Understanding mistakes and uncovering biases". In Proceedings of the European Conference on Computer Vision (ECCV), pages 498–512, 2018.

[5]. Suwannee Phitakwinai, Sansanee Auephanwiriyakul and Nipon Theera-Umpon "Multilayer Perceptron with Cuckoo Search in Water Level Prediction for Flood Forecasting". 978-1-5090-0620-5/16/$31.00 c 2016 IEEE.

[6]. Kyungmin Kim; Bichen Wu; Xiaoliang Dai; Peizhao Zhang; Zhicheng Yan; Peter Vajda; Seon Kim "Rethinking the Self-Attention in Vision Transformers", 21134734/19-25 June 2021 IEEE.