TRANFORMERS BASED MULTI-LABEL IMAGE CLASSIFICATION AND NAMING – A LITERATURE SURVEY

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

Machine learning (ML), a subset of artificial intelligence (AI), has seen a sharp rise in utilisation in recent years. Through the use of machine learning (ML), software programmes can predict results more accurately without having to be explicitly trained to do so. Speech recognition, face detection, and other applications use classification, a type of ML that divides input data into various types. Image classification is taken into account here. Multi-label classification is a technique for determining a set of labels based on the characteristics and items visible in the provided image. Transformers are categorization models that have been applied to NLP (NLP). It is now utilised in picture classification. After feature extraction, the data is passed to the transformer. It will then be able to anticipate the labels in the image. In this essay, we'll talk about Vision Transformer (ViT). It is a paradigm for multi-label image classification that can take advantage of the intricate relationships between labels and visual attributes. The Multilayer Perceptron (MLP) idea is employed in this instance to extract features from the photos. This method trains the vision-transformer to anticipate the labels that will be hidden from view based on the input that is provided. The CIFAR-10 dataset was used in this case for training. The label mask training objective is the method's essential component. During training, this model can explicitly describe label state. As a result, this model can be used to recognise wild animals, medical images, and other things. In terms of prediction accuracy, this model can operate at an efficiency of roughly 84 percent.

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

1. INTRODUCTION

The primary task in machine learning is data analysis. The most efficient method for analysing massive data and extracting its features has emerged as neural networks. There will be a number of layers in these networks for the aim of feature extraction. And the file containing these extracted feature details will be used to anticipate the future output. This is how ML operates at its core. The primary components of our suggested model are as follows.

1.1 Multilayer Perceptrons

There are numerous neural network types that have been developed as modifications to the existing networks by researchers. Multilayer Perceptron (MLP) artificial neural networks, a traditional form, are utilised in this model [5]. Each layer of neurons is connected to the others using directed graphs, so the signal path in neurons is only one way. Backpropagation is a supervised learning method used by MLP. The input is fed into one layer, processed through one or more hidden layers, and then used for various levels of abstractions inside that layer. At the output layer, also known as the visible layer, predictions are produced at the end. It may be used to learn the mapping from input to

output and is quite adaptable. The dataset's features are extracted using MLP and the transformer in the proposed model.

1.2 Vision Transformer

Convolutional Neural Network (CNN) rival Vision Transformer (ViT) [6] has arisen in recent years. CNN now holds the monopoly in computer vision, however ViT surpassed CNN in both computational accuracy and efficiency by a factor of four. Transformer models were applied in the field of natural language processing (NLP). It is now utilised in computer vision as well. In order to perform an image classification task, a transformer is applied to a series of image patches.

1.3 Why ViT is used

In real life, a single photograph may contain several items. Multilabel image classification consists mostly of visual recognition and label set prediction. Labels for the provided image include any objects, behaviours, or properties. Several features may become lost when utilising the Convolutional Neural Network (CNN) model for feature extraction after numerous encodings. However, in the case of the transformer, no data will be lost after any number of encodes because the characteristics will automatically focus on themselves. As a result, we decided to make this forecast using the ViT network.

2. LITERATURE REVIEW

2.1 General Multi-label Image Classification with Transformers

The task of predicting a set of labels relating to objects, attributes, or other entities contained in a picture is known as multi-label image classification. In this paper, we suggest An all-encompassing framework called Classification Transformer to classify images with many labels using Trans-exploiting the intricate relationships between visual labelling and features. Our strategy includes a Transformerly trained encoder to anticipate a collection of target labels with respect to a set of masked labels and visual attributes as input the output of a convolutional neural network. a crucial component of Our approach employs a label mask training goal and a to represent the state of the laser, ternary encoding technique bels throughout training as positive, negative, or unknown. Our Model displays cutting-edge performance on difficult datasets such as Visual Genome and COCO. Furthermore, our model is more versatile since it can generate convincing results for images with incomplete or additional label annotations during inference because it clearly captures the label state during training. We use the COCO, Visual Genome, News-500, and CUB picture datasets to show this new capability.

2.1 Unified Vision-Language Pre-Training for Image Captioning

Language Pre-training. Among numerous BERT variants in language pre-training. two methods that are most relevant to our approach, namely Unified LM or UniLM (Dong et al. 2019) and Multi-Task DNN.UniLM employs a shared Transformer network which is pre-trained on three language modeling objectives: unidirectional, bidirectional, and sequence-to-sequence. Each objective specifies different binary values in the self-attention mask to control what context is avail- able to the language model. MT-DNN combines multi-task training and pre-training by attaching task-specific projection heads to the BERT network. Our work is inspired by these works.

2.2 Vision-Language Pre-training.

This has become a nascent research area in the vision-language community. most similar work to ours is VideoBERT (Sun et al. 2019b), which addresses generation-based tasks (e.g., video captioning) and understanding-based tasks (e.g., action classification). However, it separates the visual encoder and the language decoder and performs pre-training only on the encoder, leaving decoder uninitialized. In contrast, we propose a unified model for both encoding and decoding and fully leverage the benefit of pre-training

2.3 Multilabel Classification

To learn the interdependency between labels for multi-label classification, approaches based on classifier chains were proposed, which capture label dependency by conditional product of probabilities. However, in addition to high computation cost when dealing with a larger number of labels, classifier chains have limited ability to capture the high order correlations between labels. With the recent progress of neural networks and deep learning, BP-MLL (Zhang and Zhou 2006) is among the first to utilize neural network architectures to solve multi-label classification. It views each output node as a binary classification task, and relies on the architecture and loss function to exploit

the dependency across labels. It was later extended by (Nam et al. 2014) with state-of-the-art learning techniques such as dropout.

2.4 Multi-Label Image Recognition

Many efforts have been dedicated to extending deep convolutional networks for multi-label image recognition. A straightforward way for multi-label recognition is to train independent binary classifiers for each class/label. However, this method does not consider the relationship among labels, and the number of predicted labels will grow exponentially as the number of categories increase. For instance, if a dataset contains 20 labels, then the number of predicted label combination could be more than 1 million. Besides, this baseline method is essentially limited by ignoring the topology structure among objects, which can be an important regularizer for the co-occurrence patterns of object.

2.5 Perceptron

Perceptron is Machine Learning algorithm for supervised learning of various binary classification tasks. Further, Perceptron is also understood as an Artificial Neuron or neural network unit that helps to detect certain input data computations in business intelligence. Perceptron model is also treated as one of the best and simplest types of Artificial Neural networks. However, it is a supervised learning algorithm of binary classifiers. Hence, we can consider it as a single-layer neural network with four main parameters, i.e., input values, weights and Bias, net sum, and an activation function. In Machine Learning, binary classifiers are defined as the function that helps in deciding whether input data can be represented as vectors of numbers and belongs to some specific class. Binary classifiers can be considered as linear classifiers. In simple words, we can understand it as a classification algorithm that can predict linear predictor function in terms of weight and feature vectors.

4. CONCLUSIONS

We suggest that an innovative and adaptable paradigm for multi-label image categorization and naming is based on transformers. It is simple to carry out this strategy. This model can discover how labels attend various portions of the image and learn simple adaptive interactions through attention. In comparison to the standard categorization methods now in use, this model performs well. We have offered a qualitative and quantitative study using this methodology. For more precise image predictions in the medical field, this model can be expanded.

5. REFERENCES

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

[2]. “Unified Vision-Language Pre-Training for Image Captioning and VQA”
Luowei Zhou,1 Hamid Palangi,2 Lei Zhang,3 Houdong Hu,4 Jason J. Corso,1 Jianfeng Gao2

[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.