Smart Intelligent Computing and Communication Technology
V.D. Ambeth Kumar et al. (Eds.)

doi:10.3233/APC210061

Intelligent Waste Classification System Using Vision Transformers

Menti Aditya Gowrish[a,1], Mukesh Kumar Dahlan[b], R. Subhashini[c]
[a1,b]*UG Scholar, Dept of CSE, Sathyabama Institute of Science and Technology, Chennai, India*
[c]*Professor, Dept of IT, Sathyabama Institute of Science and Technology, Chennai, India*

Abstract. The issue of waste management is a growing concern, ranging from agricultural fields to industries and villages to cities. Our project aims to contribute to this issue by solving the problem of segregation of waste by using the latest advances in computer vision. The Transformers for Image Recognition at Scale, which will give highly efficient results in less time compared to the models based on Convolutional Neural Networks which loses a lot of valuable information and ignores the relationship between part of images and as a whole. The Self-Attention in vision transformers gives them the capability to understand the connection between inputs, where operations can be processed in parallel on multiple GPUs and CPUs, which cannot be achieved in the case of Convolutional Neural Networks. This project allows us to build an effective system that can classify waste in real-time without human intervention in large scale industries and also in a regular household.

Keywords. Vision Transformers, Convolutional Neural Networks, Self-Attention, Image Recognition, Computer Vision.

1. Introduction

Every Year more than 3 billion ton of garbage or waste is generated all over the globe where metropolitan areas alone contribute large number to this. The amount of waste materials produced will increase by more than 70% by the year 2024. The classification of garbage using hands where people are hired for categorizing the garbage or materials. The person whoever separates the garbage material is at the risk of facing different kinds of health issues because of the toxic chemicals and substances present within the garbage. The above health issue is overcome by proposing a system called "Intelligent Waste Classification System Using Vision Transformer(Vit)". The ViT considers an input image as a series of blotch, Akin to a series of word embedding's generated by a natural language processing (NLP) Transformer. It demonstrates excellent performance when trained on sufficient data outperforming a comparable state-of- the- art CNN(Convolutional Neural Network) with four times fewer

[1]Menti Aditya Gowrish, UG Scholar, Dept of CSE, Sathyabama Institute of Science and Technology, Chennai, India.
E-mail: aditya.menti@gmail.com

computational resources. The "Intelligent Waste Classification System using Vision Transformer" management procedure of the waste materials is going to speed up and the advantage of this system does not involve any individual.

2. Motivation

The old and traditional procedure of classifying garbage or waste materials using human hand by hiring persons for separating different types of waste objects. The one who separates waste is at risk of facing various health problems due to the harmful and toxic substances present within the material or garbage. By keeping this in mind, an automated system is developed which can separate the waste materials into their respective groups. This intelligent and accurate waste classification system takes less time to segregate the waste than the physical way. The separation of garbage or waste material is going to be salvage and can be translated into different forms of power and propellant for the development of the country's wealth.

3. Objective

The proposed model "Intelligent Waste Classification System using Vision Transformer" follows an efficient procedure to segregate the waste. It classifies the garbage or waste materials such as glass objects, metal objects, paper objects, cardboard, plastic objects and trash using an intelligent waste classification system instead of a hand- picking method. The advantage of this system is that it decreases the training time significantly. It understands the connection between inputs. It handles variable sized input using stacks of self-attention instead of CNNs and RNNs.

4. Related Work

4.1 Fine-Tuning Models Comparison Garbage

UmutOzkayaeta l(2018) aimed to build a deep learning application which recognizes different kinds of trash in garbage to furnish recyclability with vision framework. Two distinct classifiers namely Support Vector Machines and Softmax were utilized to check the execution of different classifiers and six distinct kinds of images of garbage were accurately characterized with the highest accuracy of 97.86% with GoogleNet+SVM.

4.2 Intelligent Waste Classification System Using Deep Learning Convolutional Neural Network

OlugbojaAdedeji et al.(2019) developed this deep learning tool by utilizing 50-layer Residual Network pre-train CNN architecture that is a deep-learning tool as well as a feature extractor and Support Vector Machine that is employed to differentiate the waste into six classes namely cardboard, paper, glass, metal and plastic etc. The planned model was tested on the data set that was built by Gary Thung and Mindy Yang which was able to attain 87% accuracy[1].

4.3 Visual Transformers: Token-based ImagRepresentation and Processing for Computer Vision

BichenWu et al.(2020) developed this model where it represents images as semantic visual tokens and then it runs these into transformers to densely model token relationships. The Visual Transformer works in a semantic token space attending to different parts of image based on context which is in sharp contrast to pixel-space transformers that requires orders-of-magnitude and more compute. Vision Transformer using advanced training methods significantly outperform their convolutional Neural Networks raising Residual Networks accuracy on ImageNet dataset by 4.6 to 7 points while using lesser FLOPs and parameters. In the case of semantic segmentation on LIP dataset and COCO-stuff dataset, Vision Transformer based feature pyramid networks (FPN) achieves 0.35 points greater mIoU while decreasing the FPN module's FLOPs by 6.5x.

4.4 An Image is Worth 16x16 Words: Transformers for mage Recognition at Scale

Alexey Dosovitskiy at al.(2020) argued that in Computer vision, attention is used in concurrence with CNN's or used to substitute some parts of CNN's keeping its entire structure in place. They showed that this dependence on Convolutional neural networks is not required and a sequence of patches of an image is given to a pure transformer directly that performs image classification very well. Once the model is pre-trained on huge quantities of information and then when they are transferred to multiple different-sized image classification benchmarks (VTAB, CIFAR-100, ImageNet, etc.), Vision Trans- former (ViT) is able to attain marvelous results when weighed against the state of the art CNN's, all while training on considerably lesser computational resources[2].

5. Existing System

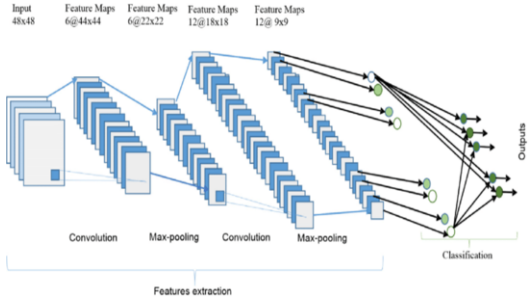
A Convolutional Neural Network (ConvNet/CNN) is a Deep learning algorithm that is able to differentiate one image from another by assigning learnable weights and biases to various features of input images. The preprocessing that is required in a Convolutional Neural Network is a considerably less weighed against other classification algorithms. At the same time as in basic methods, with sufficient training Filters are hand engineered, Convolutional Neural Networks can study these filters or traits. The Convolutional Neural Network classifies the images by analyzing an image to know if some definite parts of the image are present in that image or not. A single scalar is outputted by Neurons in ANN's except CNN's utilize convlayers that for each kernel, it replicates the same kernel's weights across the entire input volume which produces an output 2D matrix of replicated feature detector. With a part of the input volume, every number is the output of that kernel's convolution. Then, all kernel's2D matrices are stacked on top of one another to provide an outcome of a convolutional layer. Then, it is attempted to obtain perspective in-variance within the actions of neurons through the approach of maxpooling that consecutively selects the largest number in each region by looking at places inside the above defined 2D matrix. As an end result, we get invariance of activities. The invariance indicates that with the aid of converting the input a little, the same outcome still remains. The output signal of a neuron is called as an activity. In a nutshell, networks activities will no longer change

due to maxpooling even if we shift the object that we want to detect by a little bit and the network will nevertheless detect the object within the input image. The mechanism described above is not always useful, because maxpooling loses important data and does not encode relative spatial relationships between different features of the input images. Because of this, the Convolutional Neural Network is not always fixed to large transformations of input data.

5.1 Disadvantages of Existing System

- CNN fails to encode relative spatial information, but good to identify certain features in the input image but it does not consider the positioning of those features with respect to each other.

Figure 1: The overall architecture of the Convolutional Neural Network.



- CNN makes the predictions by means of analyzing an input image after which inspecting to check if some of the components that it previously analyzed are present in that image or not.
- The numbers starts at 1 with every call to the enumerate environment.
- CNN is not always invariant to large transformations of the input data.
- Parallel Processing is not possible with CNN's.

6. Proposed System

The Vision transformer System (model) is applied to attain excellent results with pure transformer architecture applied directly to a sequence of image patches for classification tasks. Additionally, it outperforms the state-of-the-art convolutional networks on several image classification tasks by using substantially fewer computational resources (at least four times fewer than SOTACNN) to pre-train. An automated classification system employing a deep learning algorithm has been used to segregate or differentiate wastes into different groups. The implementation of a well-planned system that can classify waste efficiently will increase recycling rate and reduces the burial of waste in the soil and oil pollution. The careful and robust classification of waste materials is very essential given the strict controls needed for storage, treatment and disposal of hazardous waste.

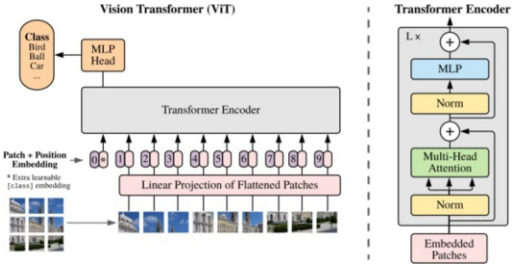


Figure 2: The overall architecture of the Vision Trans- formers.

The regular transformers use words to learn about sentences whereas Vision transformer uses pixels to achieve a similar result for images. In contrast to words, individual pixels do not convey any meaning by themselves which was one of the reasons we shifted to convolutional filters that operated upon a group of pixels. Therefore, we divide the whole image into small patches or words. All the patches are flattened using a linear projection matrix and fed into the transformer with their positions in the image. The Vision transformer model grasps to encode the distance within the input image similar to that of positional embedding. The closer patches in the input image tends to have more identical positional embedding. The column and row structure appears as patches in the same column or row will have similar positional embedding. A sinusoidal structure is sometimes apparent for larger grids.These embedded patches go through alternating layers of multi-headed self-attention, multi-layer perception (simple feed-forward neural network) and layer normalizations like in a regular transformer. Even in the lowest layers, the Self-attention in Vision Transformer allows them to amalgamate information across the entire input image. It is examined to check what degree the network makes use of this capability. A classification head is attached at the end of the transformer encoder to predict the final classes. Like any other convolutional model, one can use the pretrained encoder base and attach a custom Multi-Layer Perception(MLP) layer for fine tuning the model to suit their classification task.

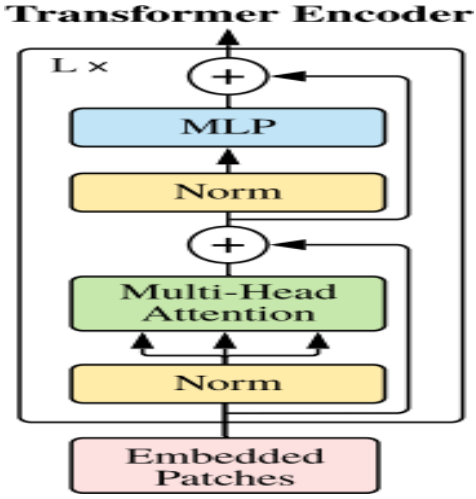


Figure 3: Architecture Diagram of Transformer Encoder.

Model	Layers	Hidden size D	MLP size	Heads	Params
ViT-Base	12	768	3072	12	86M
ViT-Large	24	1024	4096	16	307M
ViT-Huge	32	1280	5120	16	632M

Figure 4: Details of Vision Transformer model variants.

6.1 Advantages of Proposed System

- Vision Transformers encode relative spatial information. It not only detects features in the image but also considers their positioning with respect to each other.
- Self-attention allows Vision Transformer to integrate information across the entire input image even in the lowest layers.
- It uses substantially fewer computational resources.
- Self-Attention in vision transformers gives them the capability to understand the connection between inputs where operations can be processed in parallel on multiple GPUs and CPUs.
- Parallel Processing is possible with Vision Transformers.

7. Result and Discussion

The Vit model achieves about 78.89 percent accuracy and 93 percent top five accuracy on the test data after 100 epoch on the waste classification dataset. These findings are not competitive as a CNN trained from scratch on the same data can achieve 87 percent accuracy.

We can try to train the model for more epochs using a greater number of transformer layers, resize the input images, adjust the patch size or increase the projection dimensions to boost the model quality without pre-training.

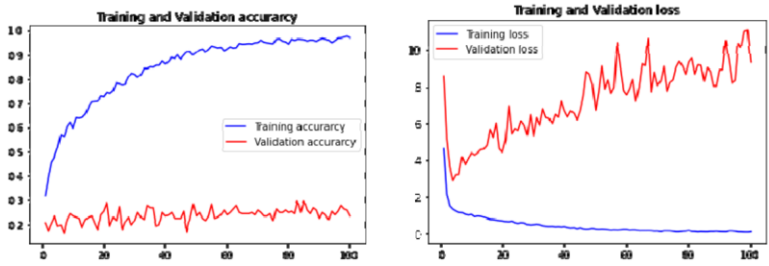


Figure 5: Training and Validation Accuracy and Training and Validation Loss

8. Conclusion

This model will make the system of waste management more efficient and flexible. The automatic classification of waste is made easier by this method without human

involvement. Consequently, it prevents contamination and different types of harmful pollution. When tested against the trash data sets, the accuracy level lasted of 78.89%. The classification procedure of the waste objects will be much quicker by using this system. When more images are provided to the dataset, then the system accuracy can be improved. In the future, more vast technology is used to improve the system to manage and classify more waste items.

References

[1] Olugboja Adedeji, ZenghuiWang Intelli- gent Waste Classification System Using Deep Learning Convolutional Neural Network, 2nd International Conference on Sustainable Materials Processing and Manu facturing, 607–612, 2019

[2] UmutÖzkaya and LeventSeyfi, "Fine-Tuning Models Comparisons on Garbage Classification for Recyclability",1(2),pp: 1-14, (2018).

[3] BichenWu, ChenfengXu, XiaoliangDai, Visual Transformers: Token-based Image Representation and Processing for Computer Vision, 5(2), pp: 1-14, (2020).

[4] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, pp:1-8, (2020).

[5] https://towardsdatascience.com/wtf-is-image- classification-8e78a8235acb

[6] https://towardsdatascience.com/end-to- end-pipeline-for-setting-up-multiclass-image- classification-for-data-scientists-2e051081d41c

[7] https://medium.com/swlh/convolutional- neural-networks-for-multiclass-image- classification-a-beginners-guide-to-6dbc09fabbd

[8] https://scikit-learn.org/stable/user_guide.html

[9] https://devdocs.io/scikit_learn/

[10] https://medium.com/analytics-vidhya/vision- transformers-bye-bye-convolutions- e929d022e4ab

[11] https://medium.com/lsc-psd/introduction-of-self-attention-layer-in-transformer

[12] R. Subha Shini et.al., " Recurrent Neural Network based Text Summarization Techniques by Word Sequence Generation",IEEE International Conference on Inventive Computation Technologies (ICICT), 2021, DOI: 10.1109/ICICT50816.2021.9358764