Vision Transformers for Image Classification using Convolution Neural Network (CNN)

Author[1] :NAGA VENKATA SAI LAKSHMI DEVI KOTTAGUNDA,
Author[2]:NELAKUDITI LAKSHMI PRASANNA SAI
Author [3]:MAJJARI MAHESHWARI
Author[4]:MOHITHA GUNDAVARAPU

[1] Student, ECE, VASIREDDY VENKATADRI INSTITUTE OF TECHNOLOGY, A.P.,INDIA
[2] Student, ECE, VASIREDDY VENKATADRI INSTITUTE OF TECHNOLOGY, A.P.,INDIA
[3] Student, ECE, VASIREDDY VENKATADRI INSTITUTE OF TECHNOLOGY, A.P.,INDIA
[4] Student, ECE, VASIREDDY VENKATADRI INSTITUTE OF TECHNOLOGY, A.P.,INDIA

ABSTRACT

Convolution Neural Networks (CNNs) are de-facto for computer vision applications. CNNs learn spatially local representations across different vision tasks based on their inductive biases. Further improvement in the performance can be achieved by learning global representations which require self-attention-based vision transformers (ViT's). However, this improvement is obtained only at the cost of ViT's being heavy-weight unlike CNNs. In this project we investigate the possibilities of building a model which is both light-weight and having low latency that are suitable for deploying on edge computing devices such as mobiles. We are planning to train and test our network on different image datasets such as Image Net that are available in public domains.

Keyword : - *Key word1*:*Image classification , Key word2:convolution neural networks, Key word3:vision transformers, Key word4 :mobile-vision transformers key word5:Inductive bias, key word 6:spatial dimensions.*

---

1. INTRODUCTION TO CONVOLUTION NEURAL NETWORKS

Deep Learning has proved to be a very powerful tool because of its ability to handle large amounts of data. CNN's were first developed and used around the 1980s. The most that a CNN could do at that time was recognize handwritten digits. It was mostly used in the postal sectors to read zip codes, pin codes, etc. The important thing to remember about any deep learning model is that it requires a large amount of data to train and also requires a lot of computing resources. This was a major drawback for CNNs at that period and hence CNNs were only limited to the postal sectors and it failed to enter the world of machine learning .The interest to use hidden layers has surpassed traditional techniques, especially in pattern recognition. One of the most popular deep neural networks is Convolutional Neural Networks.

1.1 VARIOUS TYPES OF LAYERS IN CNN

1**.CONVOLUTION LAYER**: This layer is the first layer that is used to extract the various features from the input images. In this layer, the mathematical operation of convolution is performed between the input image and a filter of a particular size MxM.

2.POOLING LAYER: The primary aim of this layer is to decrease the size of the convolved feature map to reduce the computational costs. This is performed by decreasing the connections between layers and independently operates on each feature map.

3.FULLY CONNECTED LAYER: The Fully Connected (FC) layer consists of the weights and biases along with the neurons and is used to connect the neurons between two different layers.

4.DROP OUT: Overfitting occurs when a particular model works so well on the training data causing a negative impact in the model's performance when used on a new data.

5.ACTIVATION FUNCTIONS: It adds non-linearity to the network. There are several commonly used activation functions such as the ReLU, Softmax, tanH and the Sigmoid functions. Each of these functions have a specific usage.

Convolutional neural networks are composed of multiple layers of artificial neurons. Artificial neurons, a rough imitation of their biological counter parts, are mathematical functions that calculate the weighted sum of multiple inputs and outputs an activation value. When you input an image in a CNN, each layer generates several activation functions that are passed on to the next layer. The first layer usually extracts basic features such as horizontal or diagonal edges. This output is passed on to the next layer which detects more complex features such as corners or combinational edges. As we move deeper into the network it can identify even more complex features such as objects, faces, etc. Based on the activation map of the final convolution layer, the classification layer outputs a set of confidence scores (values between 0 and 1) that specify how likely the image is to belong to a "class." For instance, if you have a CNN that detects cats, dogs, and horses, the output of the final layer is the possibility that the input image contains any of those animals

2. VISION TRANSFORMERS:

Transformer models have become the de-facto status quo in natural language processing . In computer vision research, there has recently been a rise in interest in Vision Transformers and Multilayer perceptrons.

Self-attention-based models, especially vision transformers are an alternative to convolutional neural networks (CNNs) to learn visual representations. Briefly, ViT divides an image into a sequence of non-overlapping patches and then learns interpatch representations using multi-headed self-attention in The general trend is to increase the number of parameters in ViT networks to improve the performance. However, these performance improvements come at the cost of model size and latency. Many real-world applications require visual recognition tasksto transformers . run on resource-constrained mobile devices in a timely fashion. To be effective, ViT models for such tasks should be light-weight and fast. Even if the model size of ViT models is reduced to match the resource constraints of mobile devices, their performance is significantly worse than light-weight CNNs. For instance, for a parameter budget of about 5-6 million is 3% less accurate than MobileNetv3 . Therefore, the need to design light-weight ViT models is imperative.

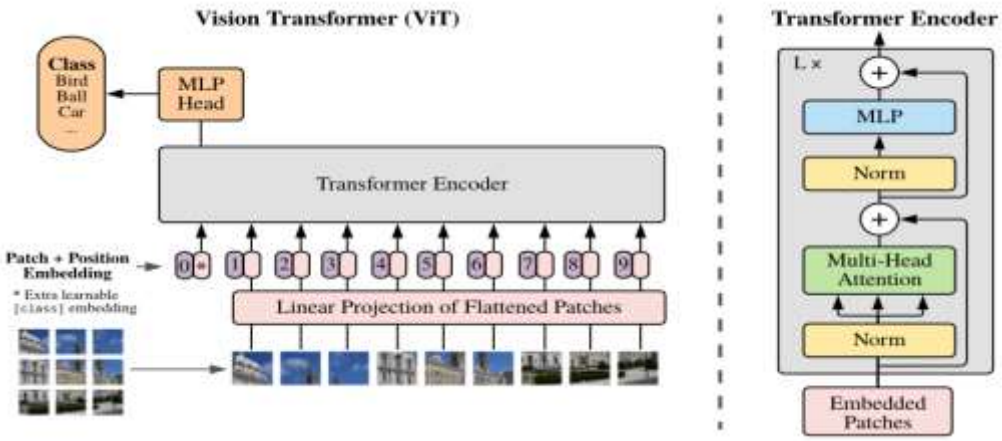

Figure -1 ARCHITECTURE OF VISION TRANSFORMERS

1. Split an image into patches of fixed sizes
2. Flatten the image patches
3. Create lower-dimensional linear em-beddings from these flattened image patches
4. Include positional em-beddings.
5. Feed the sequence as an input to a state-of-the-art transformer encoder
6. Pre-train the ViT model with image labels, which is then fully supervised on a big dataset
7. Fine-tune on the downstream dataset for image classification

2.1 TRANSFORM ENCODER

The transformer encoder includes:

- **Multi-Head Self Attention Layer (MSP):** This layer concatenates all the attention outputs linearly to the right dimensions. The many attention heads help train local and global dependencies in an image.
- **Multi-Layer Perceptron (MLP) Layer:** This layer contains a two-layer with Gaussian Error Linear Unit (GELU).
- **Layer Norm (LN):** This is added prior to each block as it does not include any new dependencies between the training images. This thereby helps improve the training time and overall performance.

Particularly, if the ViT model is trained on huge datasets that are over 14M images, it can outperform the CNNs. If not, the best option is to stick Efficient Net. The vision transformer model is trained on a huge dataset even before the process of fine-tuning. The only change is to disregard the MLP layer and add a new D times KD*K layer, where K is the number of classes of the small dataset. To fine-tune in better resolutions, the 2D representation of the pre-trained position embedding's is done. This is because the trainable liner layers model the positional embedding's.

3. MOBILE VISION TRANFORMERS

We introduce MobileViT, a light-weight and general-purpose vision transformer for mobile devices. MobileViT presents a different perspective for the global processing of information with transformers, i.e., transformers as convolutions.

To implement the MobileViT architecture which combines the benefits of Transformer and convolutions. With Transformers, we can capture long-range dependencies that result in global representations. With convolutions, we can capture spatial relationships that model locality.

Besides combining the properties of Transformers and convolutions, the authors introduce MobileViT as a general-purpose mobile-friendly backbone for different image recognition tasks. Their findings suggest that, performance-wise, MobileViT is better than other models with the same or higher complexity while being efficient on mobile devices.

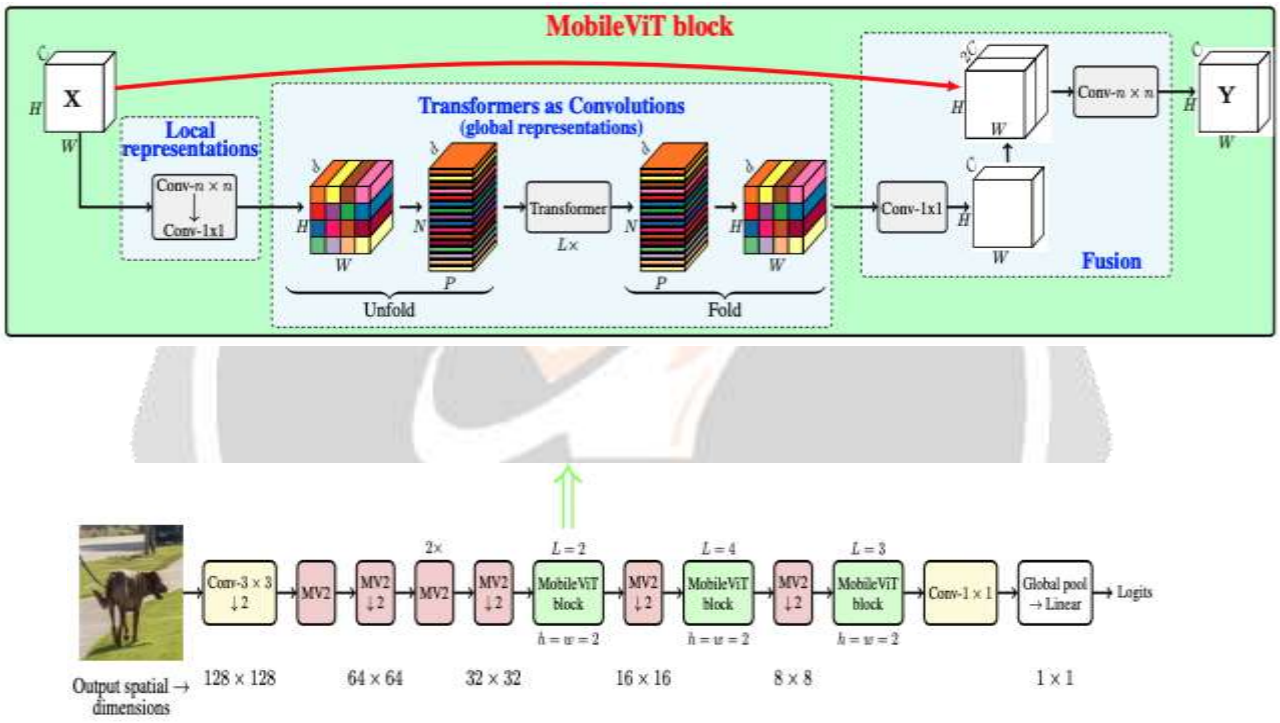

Figure 2: MobileViT. Here, Conv-n × n in the MobileViT block represents a standard n × n convolution and MV2 refers to MobileNetv2 block. Blocks that perform down-sampling are marked with ↓ 2

The MobileViT architecture is comprised of the following blocks:

- Strided 3x3 convolutions that process the input image.
- MobileNetV-style inverted residual blocks for down sampling the resolution of the intermediate feature maps.
- MobileViT blocks that combine the benefits of Transformers and convolutions. It is presented in the figure as shown above.

Advantages:

- The main advantage of CNN compared to its predecessors is that it automatically detects the important features without any human supervision. For example, given many pictures of cats and dogs it learns distinctive features for each class by itself. CNN is also computationally efficient.
- This neural network computational model uses a variation of multilayer perceptrons and contains one or more convolutional layers that can be either entirely connected or pooled. These convolutional layers create feature maps that record a region of image which is ultimately broken into rectangles and sent out for nonlinear processing.
- Very High accuracy in image recognition problems.
- Automatically detects the important features without any human supervision.

Disadvantages:

- CNN do not encode the position and orientation of object.
- Lack of ability to be spatially invariant to the input data.
- Lots of training data is required

4. CONCLUSIONS

- CNNs suffer with image specific inductive bias, on the other hand Vit's are heavy weight models unlike CNNs
- Both individually not suitable for deploying in edge computing devices such as mobile phones
- So here we obtained a method of combining them to have light weight models with high accuracy as mobile vit.

5. REFERENCES

[1]. Reference 1: rxiv.org/pdf/2110.02178.pdf

[2]. Reference 2: https://keras.io/examples/vision/mobilevit/

[3] Reference 3: https://www.upgrad.com/blog/basic-cnn-architecture/

[4] Reference 4: https://arxiv.org/pdf/2101.01169.pdf