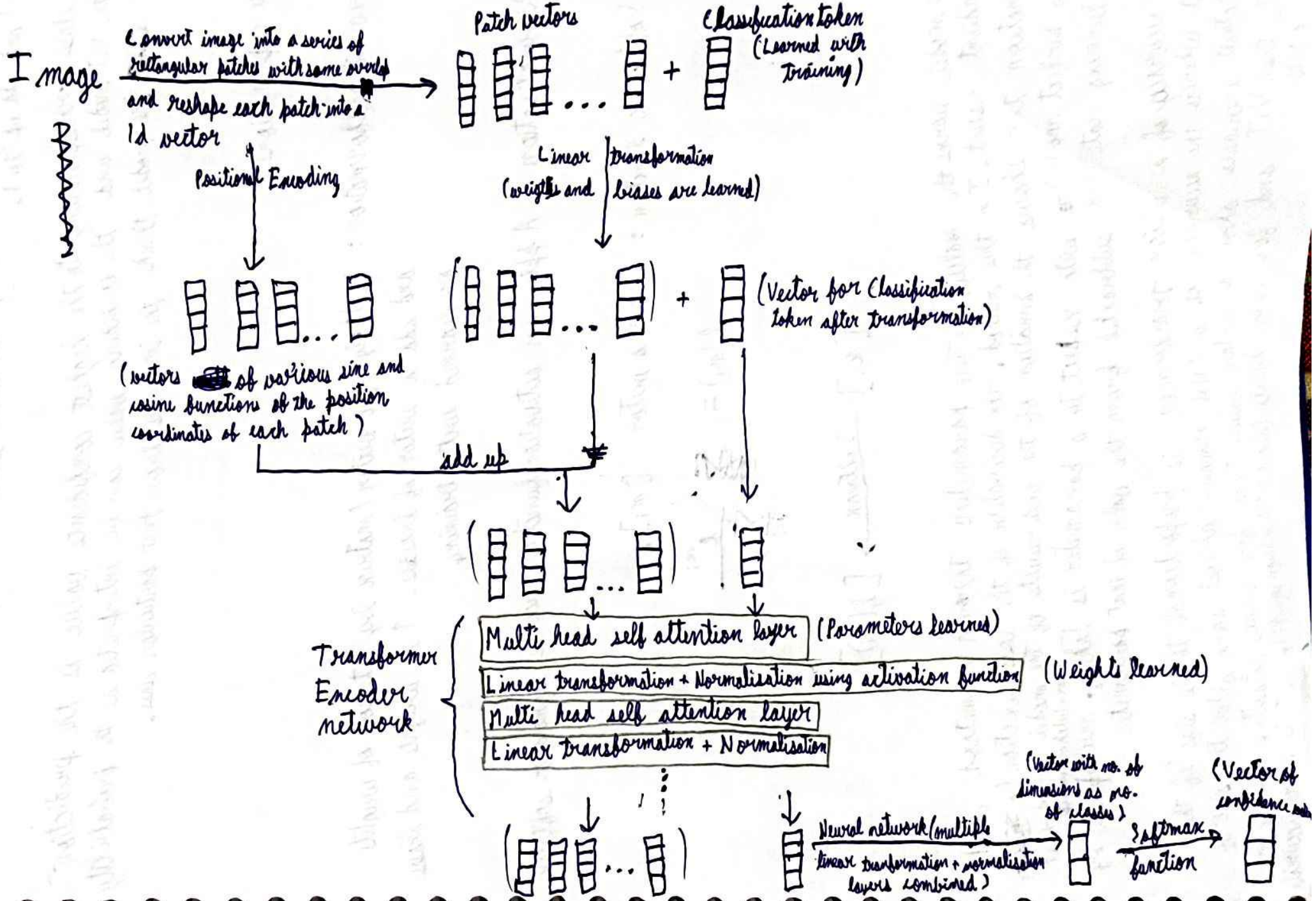
Flowchart for visual Transformers (ViT)



The vector obtained at the end is an array containing the confidence value of the image belonging to each class. The confidence values range from 0 to 1 and add up to 1.

The class corresponding to the highest confidence value is the prediction of the ViT model and the confidence value can be interpreted as the probability that the model thinks the image is from that particular class.

Some explanations:

Linear transformation: Multiply the vector / matrix by a tensor of weights and add a vector of biases. The weights and biases are learned with training

Normalisation: Apply an activation function such as sigmoid or softmax

Softmax function: For a vector $[x_i]$,

$$f(x_i) = \frac{e^{x_i}}{\sum_{j=0}^{n} e^{x_j}}$$

$$[x_i] \xrightarrow{\text{Softmax}} [f(x_i)]$$

The model learns the values of the parameters through a method called gradient descent. In this method, the derivative of the loss function (a function that shows the deviation of the end results of the model from that of a perfect model) with respect to a parameter is taken for a batch of examples, multiplied by the learning rate and subtracted from the value of that parameter

The accuracy of a vision transformer is proportional to the size of the dataset whereas the accuracy of a CNN remains almost same after the size of the dataset increases above a certain value. Thus for small datasets, CNN is more accurate than ViT and for large datasets (above ≈ 300 million images), ViT is more accurate than CNN.

Multi-head attention layers