for visual Transformers (VIT) Patch vectors Classification token Convert image into a series of (Learned with Training) rectangular patches with some overly 1 mage and reshape each patch into a ld vector Linear transformation Positional Emoding (weight and biases are learned) (Vector for Classification John after transformation) (witers at partient sine and essine bunction of the position coordinates of each patch) add up Multi head self attention layer (Parameters leavines) Transformer (Weights leavened) Linear transformation + Normalization using activation function Emoder nulli head self attention layer network Linear transformation + Normalisation ( vector with mo. of (Vector of dimensions as moundilence me Neural network (multiple restablishment materialisates result layou combined)

The victor obtained at the end is an evergay containing the confidence value of the image belonging to each class. The confidence values range from Oto I and all up to I.

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

Some emplanations:

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

Normalisation: Apply an activation function such as sigmoid or softman

Softmax function: For a vector [n;],

$$f(n_i) = \sum_{j=0}^{\infty} \frac{\ell^{n_i}}{\sum_{j=0}^{\infty} \ell^{n_j}}$$

$$[n:]$$
  $\xrightarrow{Softman}$   $[f(n:)]$ 

The model leavens 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 brom that of a perfect model) with respect to a parameter is taken, multiplied by the leaving rate and subtracted from the value of that parameter

the accuracy of a vision transformer is proportional to the size of the detect whereas the accuracy of a CNN rumains almost same after the size of the the dataset invesses above a certain value. Thus for small datasets, ENN is more accurate than ViT and for large lasets (above ~ 300 million image), ViT is more accurate than CNN.

