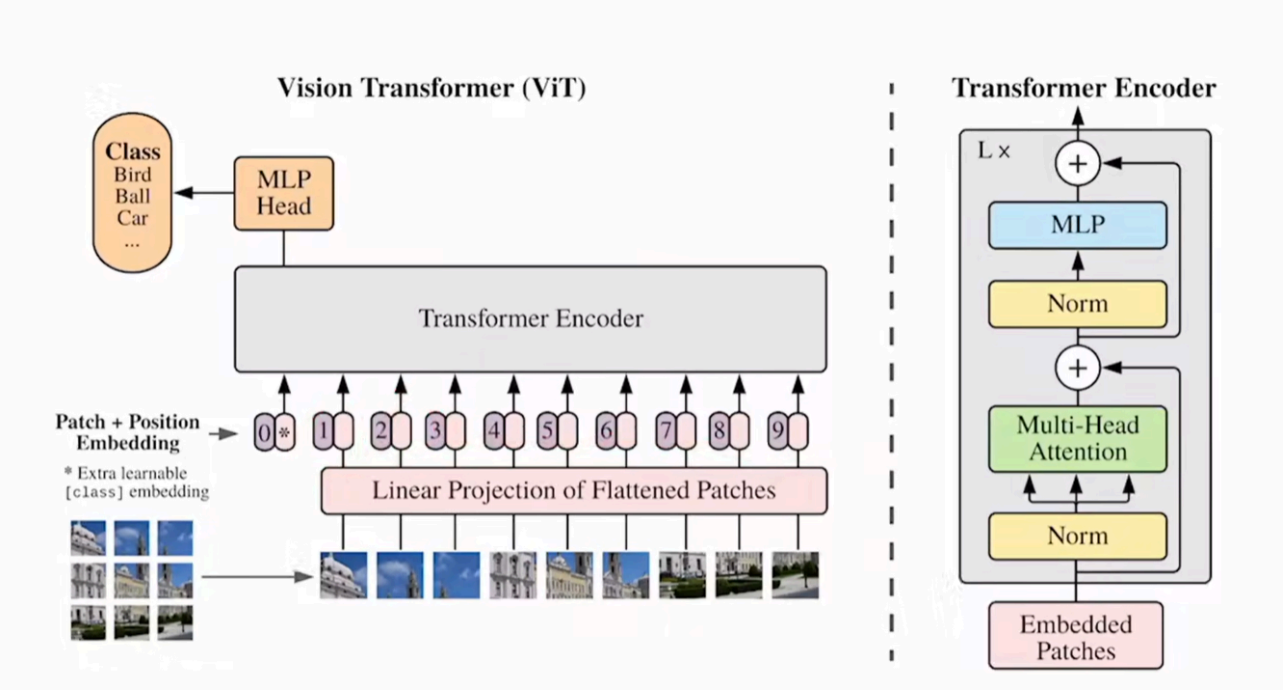
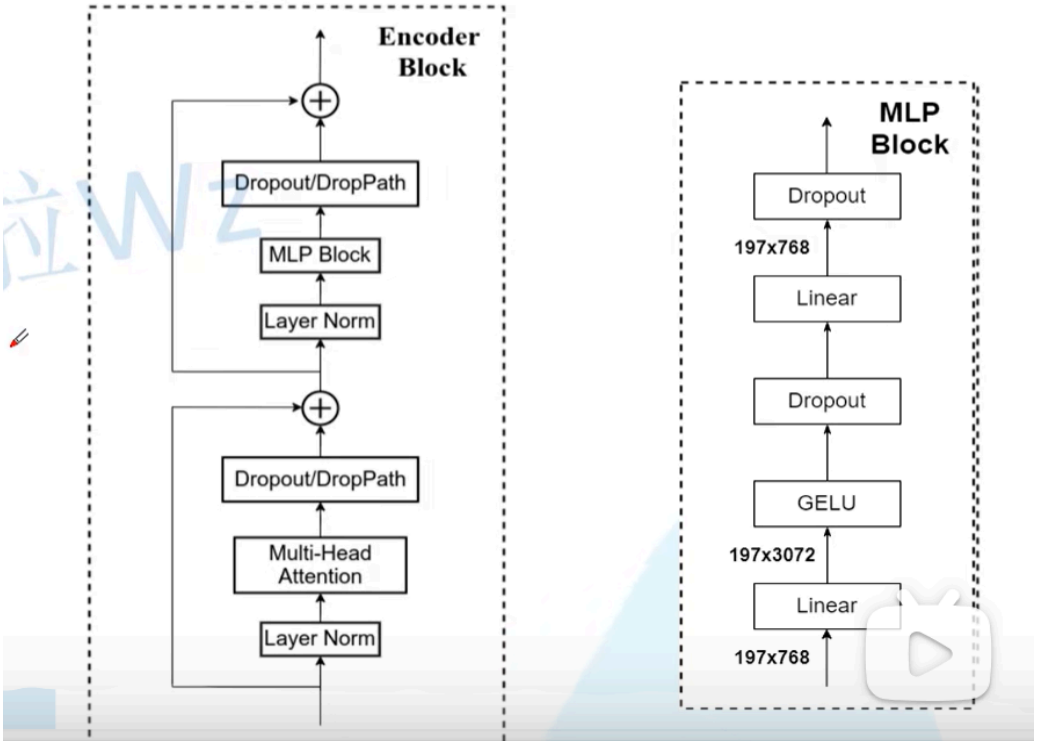
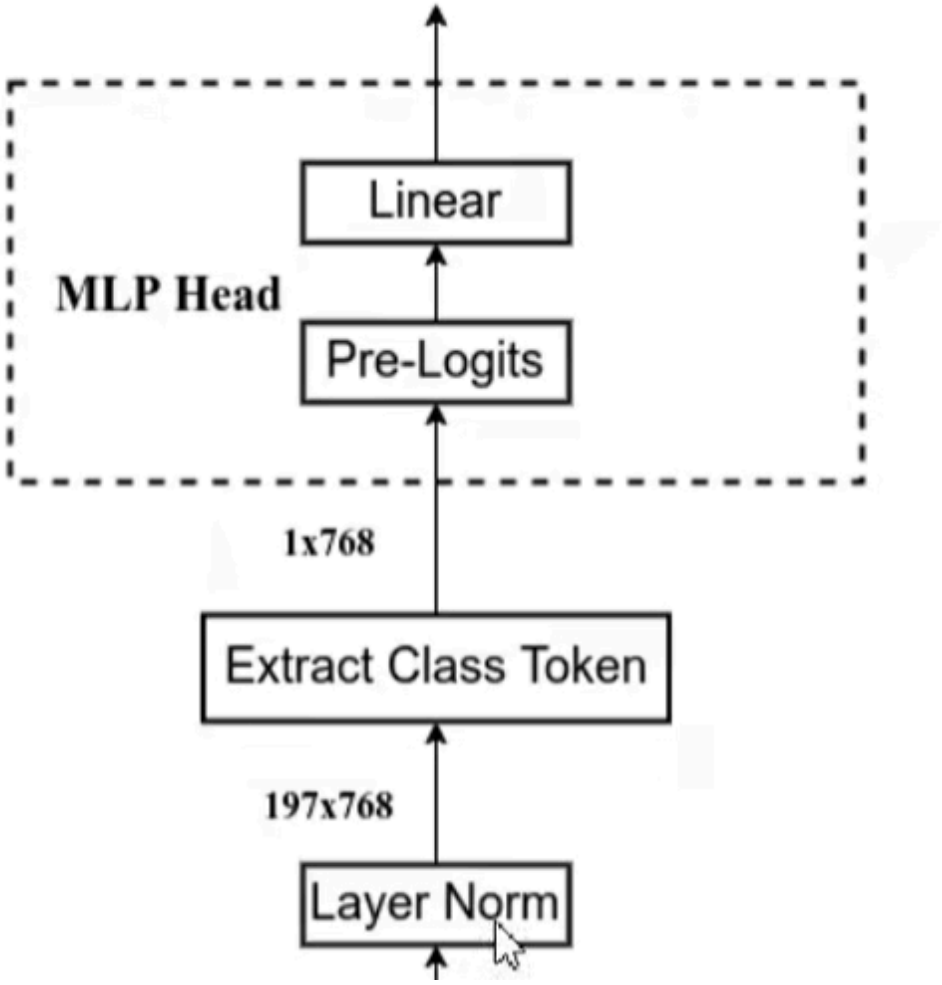
Vision Transformer



- Embedding层：transformer的input：[num_token,dim_token]
  - 将图片拆成很14*14个size为16*16的patch 然后当作input，flatten之后应该是14*14个长度为16*16*3的vector
    compute：[224*224*3]->[14*14,3*16*16]->[196,768]
  - 通过卷积核实现(kernel_size=16*16,stride=16,padding=0,num_kernel=768)
  - 拼接class[token] Cat([1,768],[196,768])=[197,768]
  - 叠加pos embedding [197,768]->[197,768]
- Transformer Encoder层：
  - 论文中Encoder Block重复L=12次



MLP Head层



- encoder完之后layerNorm一下 然后取切片 取我们的class token来进行输入class token记录类别信息，而其他的token记录的是图片信息
- 分情况ImageNet21k 用linear+tanh激活+linear
- ImageNet1k 用一层linear就行了
- 后面给一层softmax