1. Consider the T5-Large pretrained model, which has a model-dimensionality of 1024, and a vocabulary size of 32k. It has 24 transformer layers, and has 16 attention heads (for all different attentions). The feed-forward network has 4096 nodes. How many parameters does T5-Large has?

Ans. $12 d^2 * L$
$+ 16 d^2 * L = 705m$ (approx)

Don't consider - embedding matrix
- unembedding matrix
- position embedding

2. Suppose you are using (Vision) Transformer to encode images of dimensions 224×224 using patches of size 8×8. Assume that the model dimensions are 512, and a learnable class token embedding is prepended. What will be the number of learnable parameters for the (a) positional embeddings, and (b) input representation?

Ans # patches $= \frac{224 \times 224}{8 \times 8} = 784$

learnable p.e. $= (784+1) \times 512 = 401,920$

i/p $= (8 \times 8 \times 3) \times 512 = 36,864$