Data Efficient Image Transformer (DeiT)

Problems with ViT:

* Trained on a massive dataset - JFT-300M (300 million dataset)
* Huge computation - 600M parameters
* Poor generalization on small dataset.

Goal of DeiT:

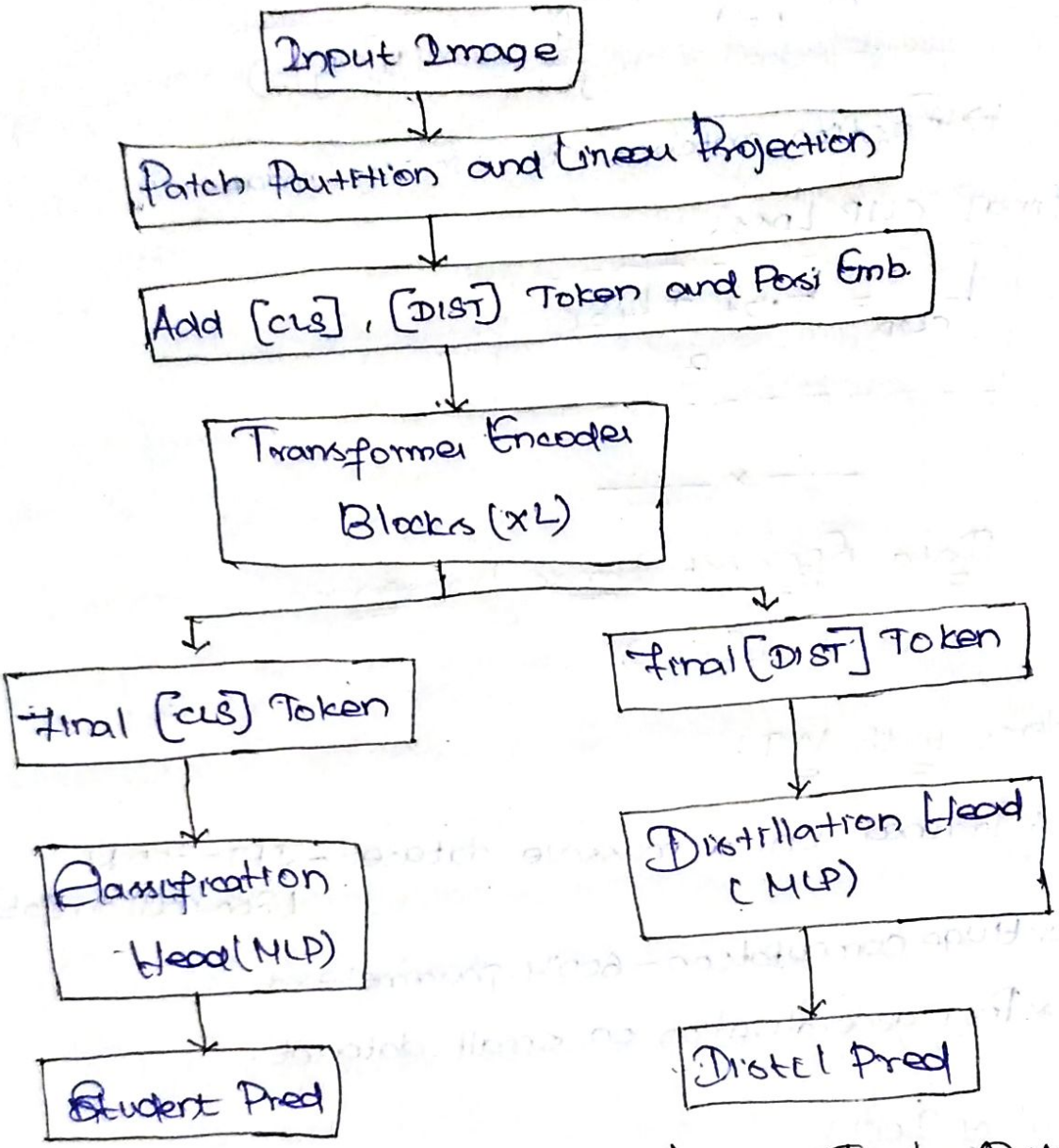
* Train a Vision Transformer efficiently using limited data and reasonable compute, without sacrificing accuracy.

→ use ImageNet-1K dataset (~1.2M Images)

→ Use smaller models (~84M params)

→ achieve performance comparable to SOTA CNNs.

Block Diagram:



⇒ The teacher output comes from a seperate, pre-trai teacher model (often a CNN like Resnet)

*Dect = Vit + one additional token...

- CLS token learns from ground truth.
- DIST token learns from teacher.

Classification Loss: (CLS Token)

$$L_{CE} = -\sum_i y_i \log p_i^{cls}$$

y_i → ground truth

p_i^{cls} → CLS Token.

Distillation Loss:

$$L_{KD} = T^2 \sum P_i^{teacher} \log\left(\frac{P_i^{teacher}}{P_i^{dist}}\right)$$

* T is used to control the softmax sharpness before computing the loss.

* Extra T^2 is introduced to cancel (γ_{T^2}) while computing backprop.

Final Loss:

$$L = \alpha L_{CE} + (1-\alpha) L_{KD} \; ; \text{Typically } \alpha = 0.5$$