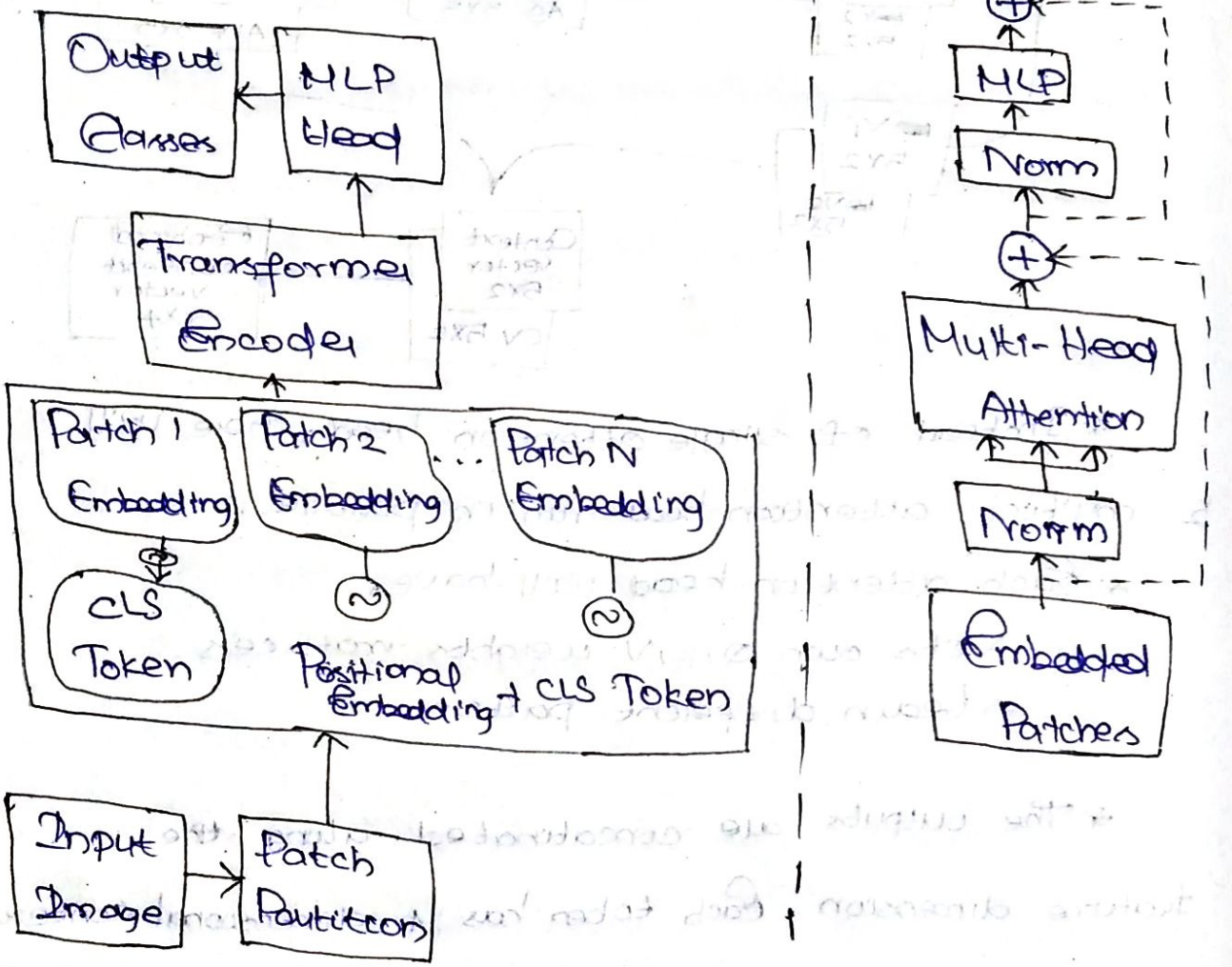
Vision Transformer

Image → Tokenization → Self attention without masking → Uses context vector of zeroth [CLASS] token for classification.

ViT Architecture:



* It converts image into sequence of tokens, then process it exactly like text using a Transformer encoder.

* No decoder | causal masking. Pure encoder-based architecture.

ViT Pipeline:

1) Images → Patches
2) Patches → vectors (tokens)
3) Add class token
4) Add positional embedding
5) Pass tokens through Transformer encoders.
6) Use class token output for classification.

—— x ——