→ Góc nhìn ≠ về Bài toán Object Classification

↓

Sự kết hợp giữa CV và NLP

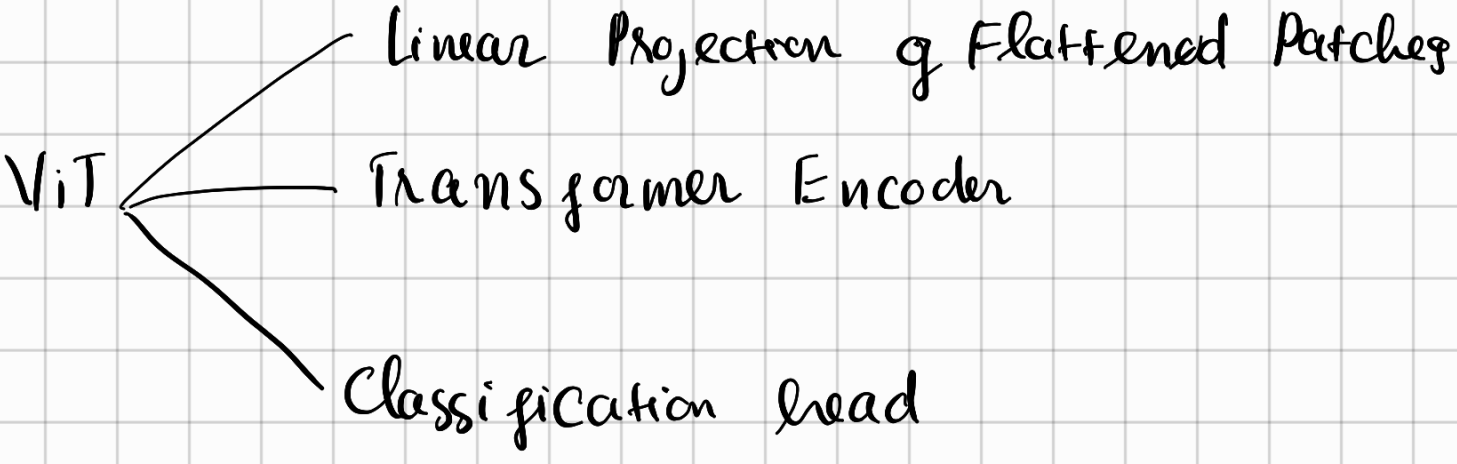
↓

Visision Transformer (ViT)

↓

Là sự kết hợp giữa 1 phần kiến trúc của

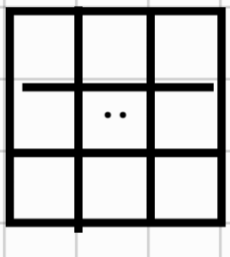
Transformer và các khối MLP
(Multilayer Perceptron)



① Linear Projection of Flattened Patches

2.1 Patch Embedding

↳ chia ảnh ra làm các patch nhỏ

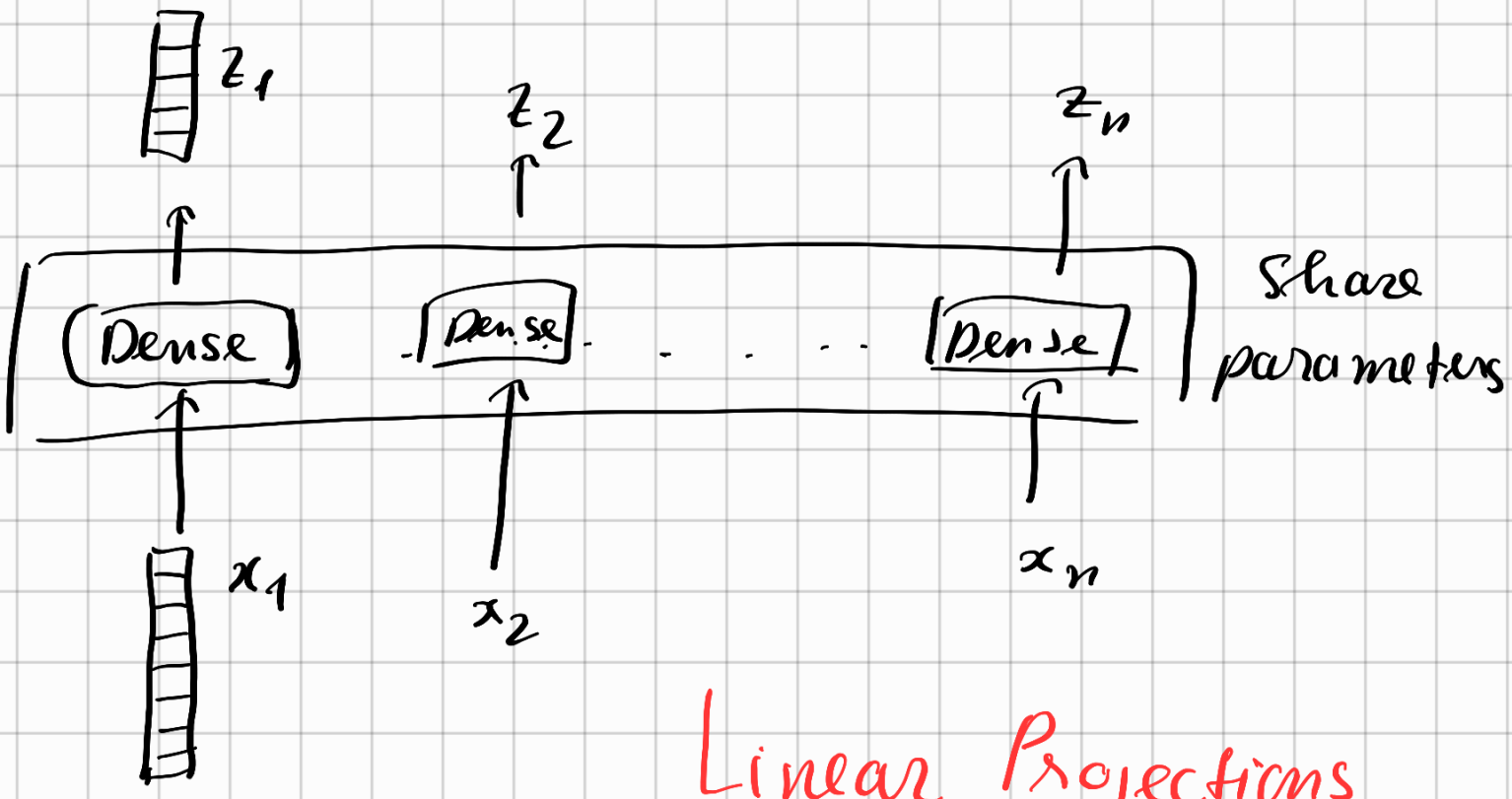


split → 9 batch.

image

↳ Đưa cái patch về dạng vector = flattend các patchs ra

- như hình



Linear Projections

↳ bản chất: là một lớp Dense vs input là flattend vector của các batch

→ output: embeeding vector tương ứng vs từng batch

$z_i = W * x_i + b$ | x_i: flatted vector
z_i: output of x_i
W: matrix embedding

2.27 Positional Embeedding

thêm thông tin về vị trí cho mỗi patch ở đầu z_i

2.3T Class Embedding

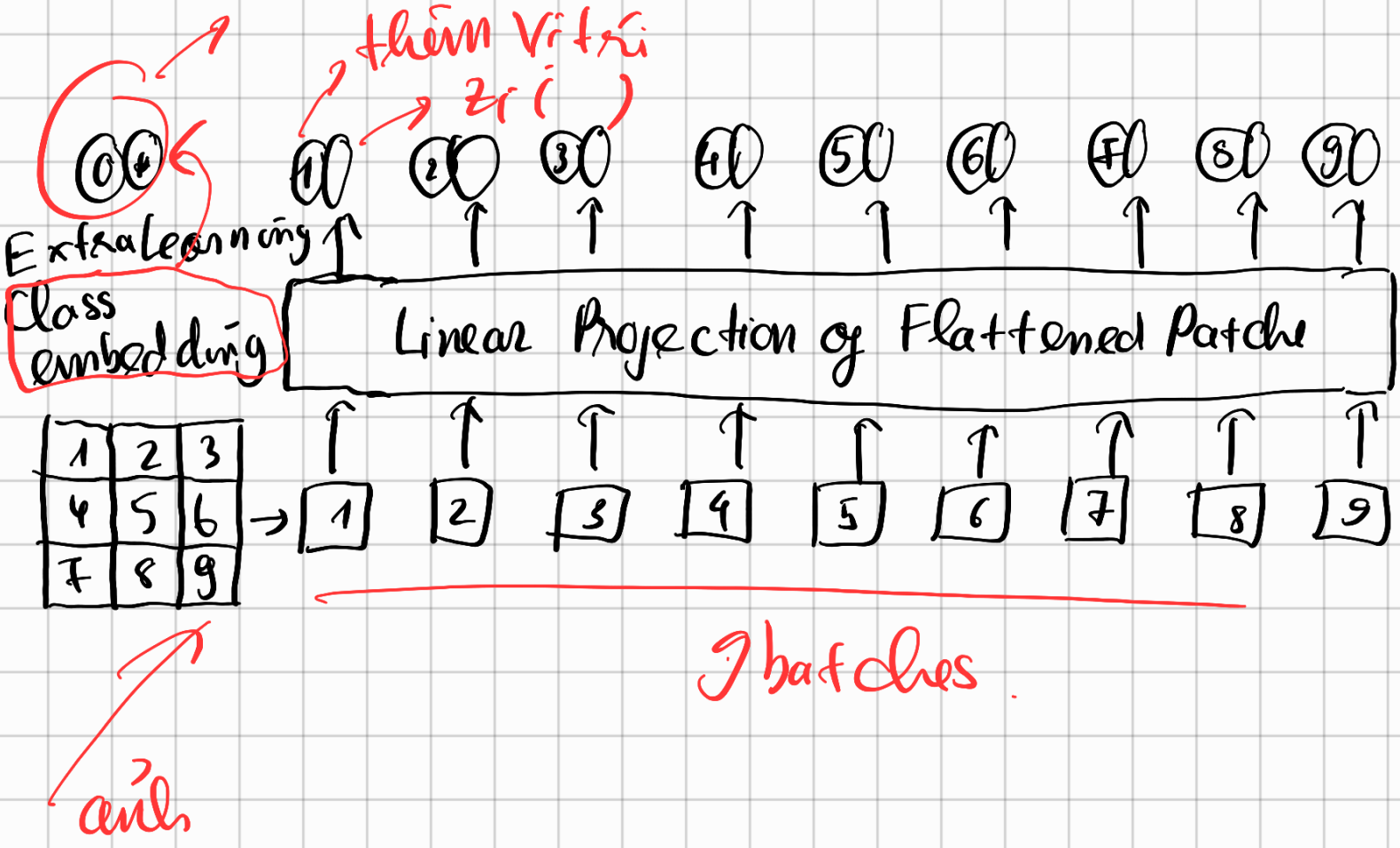
→ ý tg là mã hoá class

→ giúp mô hình "hiểu" hơn về các out put class label.

⊕ Tổng quan :

Sau step này từ một bức ảnh giả sử ảnh được chia là 3x3

mô hình ở dưới nhé



2-2? Transform Encoder

+? Self Attention layer

+? Multi head Attention

2.3? Classification Head.

một khối MLP (Multilayer Perceptron)

Input context vector c

Sxuất tương ứng vs các class

37 Training Strategies

3 bước training

+? Khởi tạo Pre training

+? Fine - tune

+? Testing