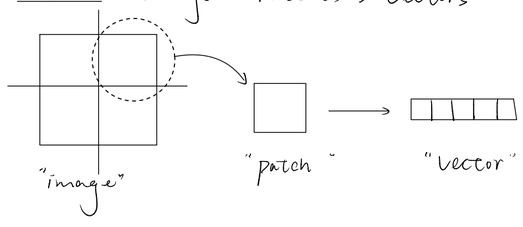
ViT: Vision Transformer

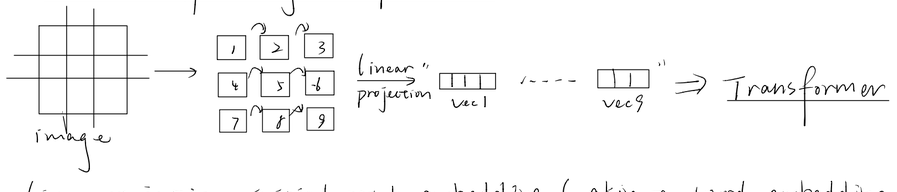
image: 3D tensor.

Key Question: How to apply Self-Attention to images?

Answer: Image ⇒ Patches ⇒ Vectors




Order: Left to right, top down

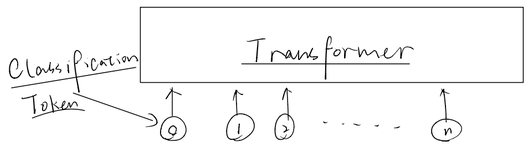



⇒ Transformer

Linear projection: initial patch embedding ( akin to word embedding )

Position embedding: Learned or manully set ( L→R, T→B )

An extra "Classification Token" is added




Input X: H × W × C

patch Size: P × P.

⇒ # of patches $\frac{H \times W}{P^2}$ ⇒

flatten vector: $1 \times P^2 \times C$ multiplied with $E \in \mathbb{R}^{(P^2 \times C) \times D}$

⇒ D-dimensional Vector.

$$z_0 = \begin{bmatrix} x_{cls} E \\ x_p^1 E \\ \vdots \\ x_p^N E \end{bmatrix} + E_{pos} \quad (E, E_{pos} \text{ are learnable})$$

$z_l' = MSA(LN(z_{l-1})) + z_{l-1}$ Residual connection

- MSA: Multi-Head Self-Attention.
- LN: Layer-Normalization

$z_l = MLP(LN(z_l')) + z_l'$ $(l \in 1, 2 \dots, L)$

MLP: Contains two layers with GELU non-linearity

ViT has less Image-Specific inductive bias (than CNN)

CNN:
- locality
- two-dimensional neighbor structure
- translation equivariance

(Inductive Bias)

ViT: MLP layer

MSA layer ⇒ Global

⇒ Weak Assumption ⟶ Need more data

Small dataset ⇒ CNN better

Large dataset ⇒ ViT better

Learning Pattern of ViT

recover the grid structure of original image

- Lower layer ⟶ local / global feature (Both)
- higher layer ⟶ global feature (only)

- ViT features are more interpretable. (more meaningful)
- ViT is more robust to noise & domain shift

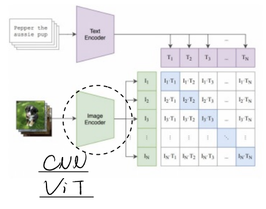
CLIP: Contrastive Language-Image Pre-training

training on image-text pairs

⇒ Vision-Language Model (VLM)

Contrastive Loss function

$$\mathcal{L}_{CLIP} = -\frac{1}{2N}\sum_{i=1}^{N}\log \underbrace{\frac{\exp(I_i \cdot T_i)}{\sum_{j=1}^{N}\exp(I_i \cdot T_j)}}_{P(T_i|I_i)} - \frac{1}{2N}\sum_{i=1}^{N}\log \underbrace{\frac{\exp(I_j \cdot T_j)}{\sum_{j=1}^{N}\exp(I_j \cdot T_i)}}_{P(I_j|T_j)}$$




- for each image $I_i$ the paired text $T_i$ should have the maximum probability
- for each text $T_i$ the pairedimage $I_i$ should have the maximum probability

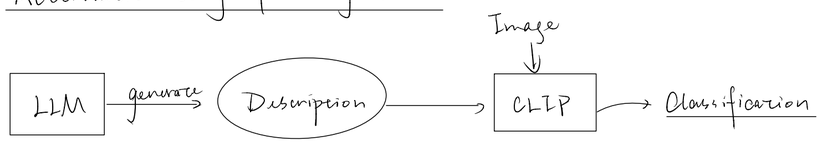
Image feature ⟶ Classification Head

Zero-Shot CLIP shows Competitive performance

Linear Probing with CLIP outperform the best ImageNet Model

CLIP is more robust to distributional shift

Alternative way of using CLIP:




CLIP Scaling with larger training dataset