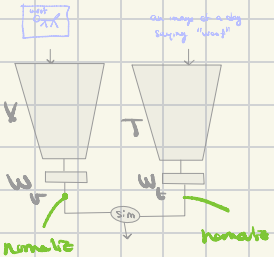
CLIP objective

Let X and Y be an image and text matrix, respectively, such that $\{(x_i, y_i)\}_{i=1}^{N}$ are image-caption pairs, where

$x_i \in \mathbb{R}^{(C, H, W)}$ and $y_i \in \mathbb{R}^{L}$

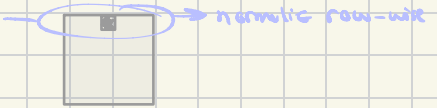


Info NCE / contrastive Loss

$$e_i = W_v V(x_i) \rightarrow \hat{e}_i = \frac{e_i}{\|e_i\|}$$

$$w_i = W_t T(y_i) \rightarrow \hat{w}_i = \frac{w_i}{\|w_i\|}$$

Let $S_{ij} = \cos(\hat{e}_i^{img}, \hat{w}_j^{txt}) * 1/\tau$, be a similarity matrix for a batch (B = batch-size) of image-caption pairs, where τ is a learnable temperature



$$p(j|i) = \frac{S_{ij}}{\sum_{k=1}^{B} \exp(S_{ik})}$$

$$q(i|j) = \frac{S_{ij}}{\sum_{k=1}^{B} \exp(S_{kj})}$$

$$p(j|i) = \frac{S_{ij}}{\sum_{k=1}^{B} \exp(S_{ik})}$$

normalize row-wise

$$q(i|j) = \frac{S_{ij}}{\sum_{k=1}^{B} \exp(S_{kj})}$$

normalize column-wise

$$\mathcal{L}_{i2t} = \frac{1}{B}\sum_{k}^{B} \log p(k|k)$$ (img, txt)

normalize

$$\mathcal{L}_{t2i} = \frac{1}{B}\sum_{k}^{B} \log q(k|k)$$ (img, txt)

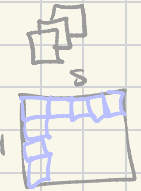
normalize

$$\mathcal{L}_{CLIP} = \left(\mathcal{L}_{i2t} + \mathcal{L}_{t2i}\right)/2$$

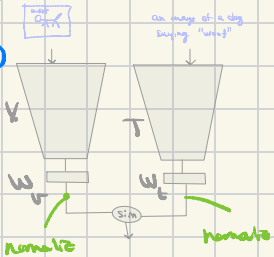
Patch Embedding

(Broadcast variable)

CLS token $(1, 1, D)$



we want (B, P, D)



ps = patch size

Conv2D (in_ch = 3, out = D, kernel size = ps, stride = ps)

so now $P = H/ps * W/ps$

Patch Embedding

x

↓

Conv2D $(B, D, H/p, W/p)$

↓

Flatten (2) (B, D, P) $P = H/p * W/p$

↓

transpose (2, 1) (B, P, D)

↓

cls_token $(B, P+1, D)$

↓

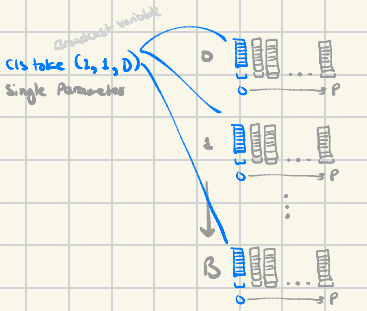
pos_embed

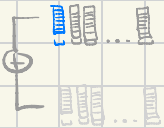
↓

x

cls token $(1, 1, D)$ — Broadcast variable

Single Parameter





Attention Head

x

↓

Q, K, V

$A = mask\left(\frac{QK^T}{\sqrt{n}}\right)$

n = head size = D / n_heads

↓

$A = Softmax(A)$

↓

$A = AV$

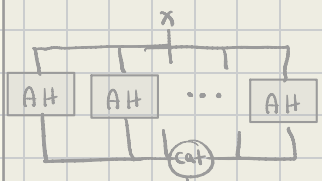
↓

Dropout

↓

x

Multi Head Attention



↓

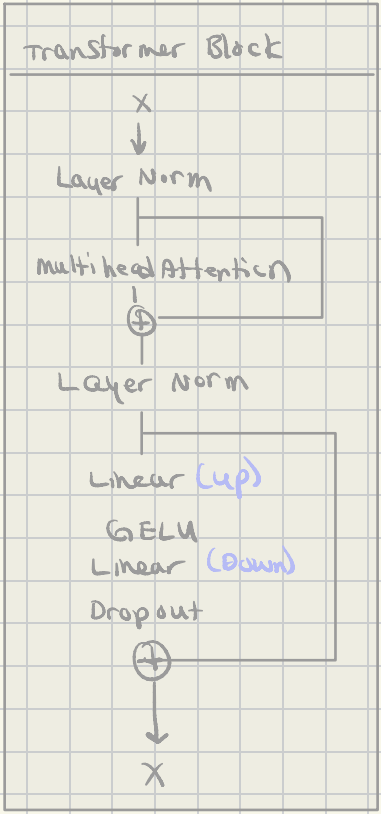
project

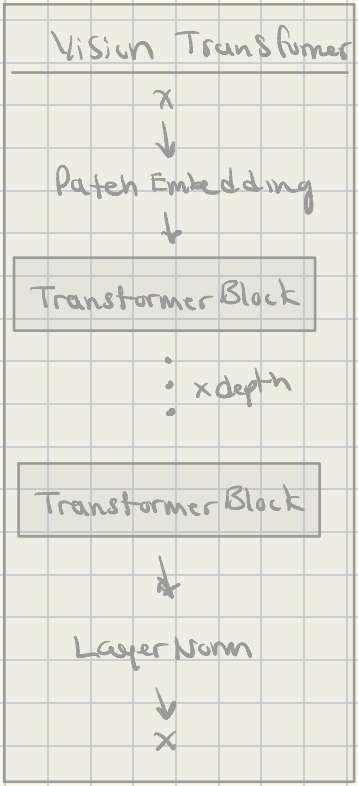
↓

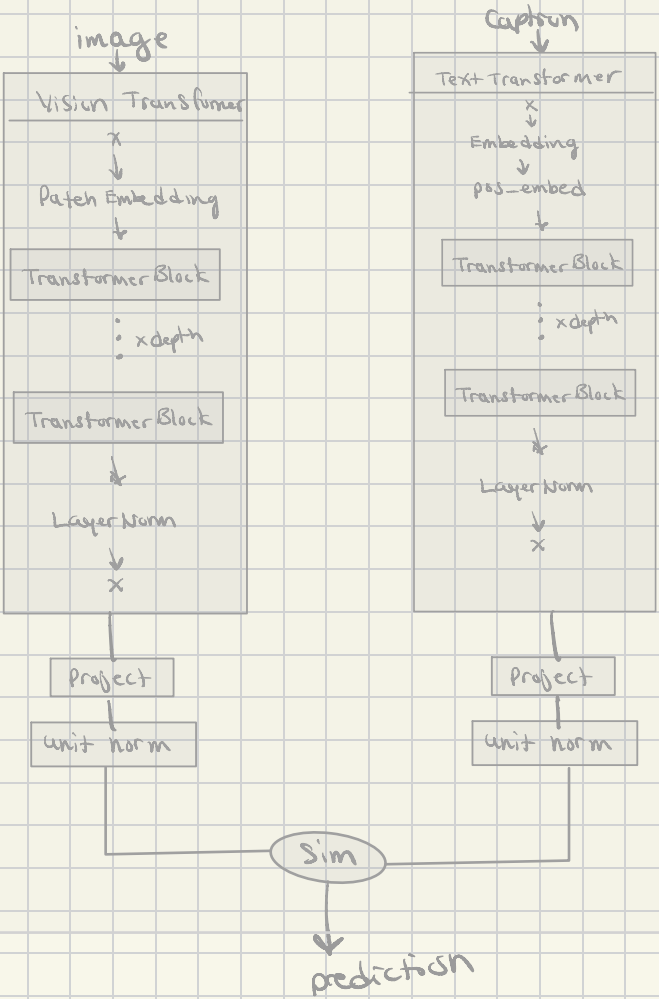
Dropout

↓

x


Transformer Block
x
Layer Norm
Multi Head Attention
Layer Norm
Linear (up)
GELU
Linear (Down)
Dropout
x


Vision Transformer
x
Patch Embedding
Transformer Block
x depth
Transformer Block
Layer Norm
x


CLIP
image
Vision Transformer
Patch Embedding
Transformer Block
x depth
Transformer Block
Layer Norm
Project
unit norm
Caption
Text Transformer
Embedding
pos_embed
Transformer Block
x depth
Transformer Block
Layer Norm
Project
unit norm
Sim
prediction