VIT

- $\frac{H \cdot W}{p \cdot sz \times p \cdot sz} \times \underset{\text{in channels}}{3}$

- flatten

Pixel 1 — R
G
B
⋮
Pixel last — R
G
B
of path

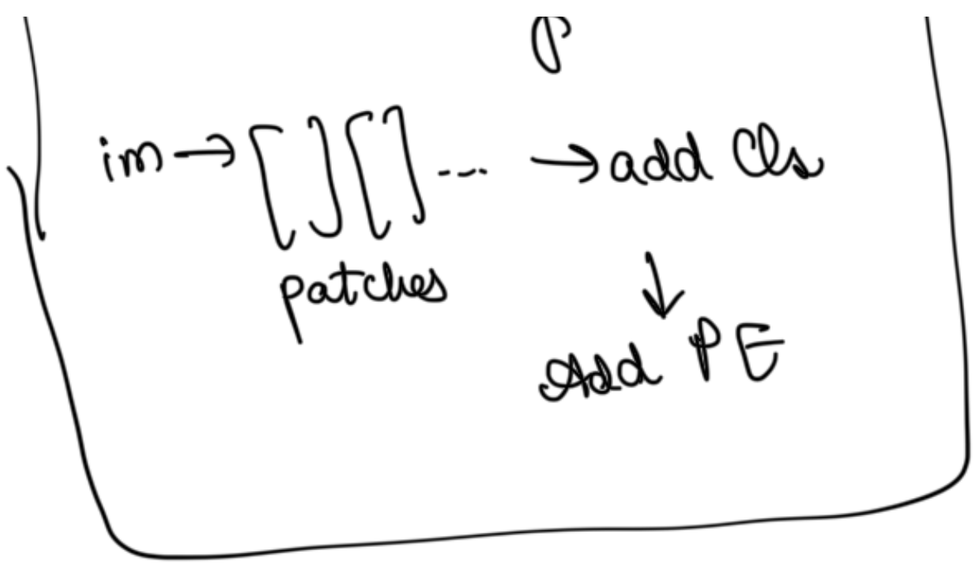
→ FC ⟶ #patches x D

#patches x channels
x patch ht x patch width

RGB patch ⟶ D dimension → Position Encoding → CLS token at pos 0

for entire image

Patch Embedding



Enops

eg. $\frac{224 \cdot 224}{16 \cdot 16} = 196$ patches

- 16×16 words

Requirement : Find relevance b/w Patch1 and Patch 4

Existing representation might not cater to whats relevant

WQ

WK

WV

learn

- Criteria of Relevance
- how to so that criteria transform patch based on similar patches

give similar products

D x head dimension

$(1 \times D) \times (D \times \text{head dim})$

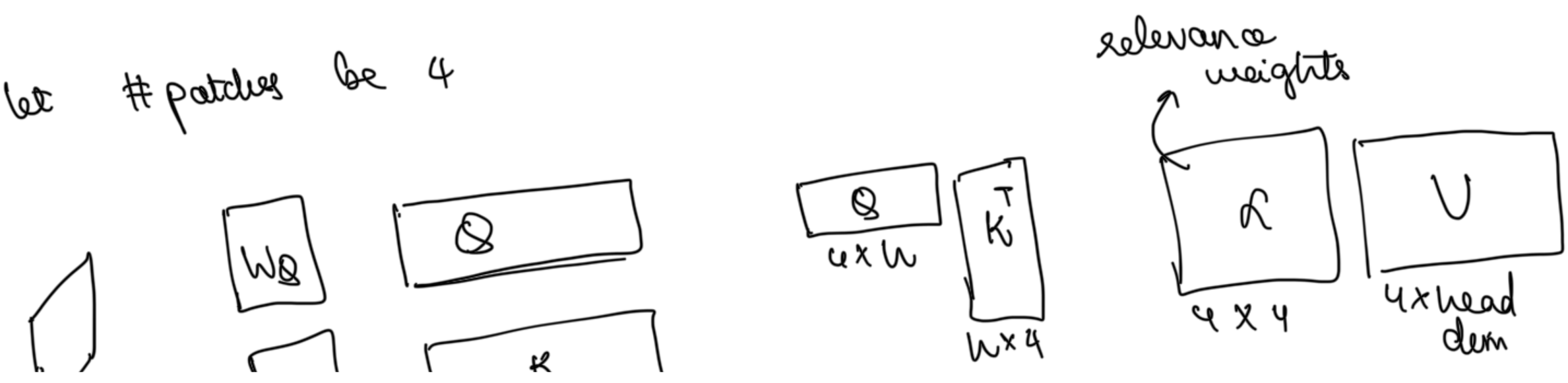
↑ Patches will be transformed to this dimension

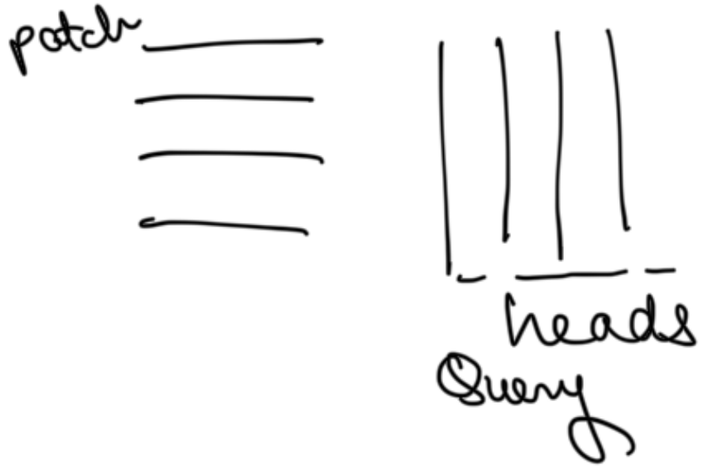
$$\alpha_i = \frac{Q K_i^T}{\sqrt{D_h}}$$

for patch 1, use Q1 to find which patches are most relevant to it

$(\alpha_1 \quad \alpha_2 \quad \cdots \quad \alpha_{\#patches}) \leftarrow$ softmax

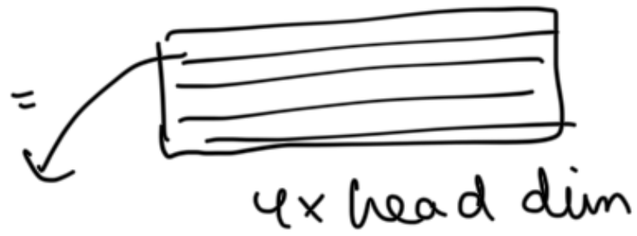
let #patches be 4





$q_{i,j}$

Query i Key j



context representation of patch i

→ Relevance for multiple factors

→ multiple heads

4 x #heads x head dim → concatenate → fc layer → W → 4 x D

So one W is enough, split to unback into Q, V, K

#p xD

D x #heads x 3 x #dim$_{head}$

→ #patches · #heads · 3 · #dim$_{head}$

Split into three

↓

Split each head & rearrange

Q, K, V,

↓

Scaled dot product and softmax

↓

Compute weighted V

↓

Rearrange, join each head

↓

project onto D

Self Attention Block

↳