Qwen2-VL Vision Encoder (Qwen2VisionTransformerPretrainedModel)

Input Image / Video

$(B, 3, T, H, W)$

PatchEmbed

Conv3d$(3 \rightarrow 1152, \ k=[2, 14, 14], \ s=[2, 14, 14])$

reshape $\rightarrow$ view$(-1, \ 1152)$

$(N, \ 1152)$

Qwen2VLVisionBlock $\times 32$

residual

LayerNorm (norm1)

apply to Q,K

VisionAttention

QKV Linear$(1152 \rightarrow 3456) \rightarrow$ Window Attn $\rightarrow$ Proj

window attention
via cu_seqlens

$+$

LayerNorm (norm2)

residual

VisionMlp

Linear$(1152 \rightarrow 4608) \rightarrow$ QuickGELU $\rightarrow$ Linear$(4608 \rightarrow 1152)$

$+$

$(N, \ 1152)$

2D RoPE

VisionRotaryEmbedding
head_dim // 2 per axis
(height, width) positions

Legend

Attention

MLP

LayerNorm

PatchMerger

2D RoPE

Residual

PatchMerger

LayerNorm (ln_q)

Reshape: view$(-1, \ 1152 \times 4)$ [2×2 spatial merge]

Linear$(4608 \rightarrow 4608) \rightarrow$ GELU $\rightarrow$ Linear$(4608 \rightarrow d_{\text{LLM}})$

$(N/4, \ d_{\text{LLM}})$

To LLM (Qwen2-VL Language Model)