11/oct/24

QWEN 2 VL

[Paper Submitted on 18 Sept 24]

→ SOTA image Understanding – Various resolutions, aspect ratios

→ Extended Video comprehension – over 20 min long videos,
high quality video based Q&A,
content creation, dialog

→ Can integrate with mobiles

→ Apache licence – 2B, 7B

→ API → 72B

→ Image Understanding :- advanced object recognition,
hand written text recognition,
multiple languages,
mathematical problem Understanding
and solving,
Chart understanding, highly distorted
document / Image Understanding,

→ Limitation :- knowledge till June 2023,
Cant extract audio from video,
Complex instruction and scenarios,
counting, Character recognition, 3D Spatial awareness

→ Model Architecture :-

[ViT] Vision Transformer + QWEN 2 Language Model

600M parameters

Image + video

① Added "Naive Dynamic Resolution" Support

↓

Arbitary Image Resolution and Dynamically mapping them into various number of visual tokens.

② Added " Multimodal Rotary Position Embeddings (M-ROPE)"

deconstruct Original Rotary Embedding to 3 Categories

3D

2D

1D

helps simultaniously capture these 3 positional information to understand complex multimodal data