

11/oct/24

QWEN2 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 QA, content creation, dialog

→ Can integrate with mobiles

→ Apache license - 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,
Can't extract audio from video,
Complex interaction and scenarios,
counting, Character recognition, 3D Spatial awareness

→ Model Architecture :-

(ViT) Vision Transformer + QWEN 2 Language Model

600M parameters

Image + Video

① Added "Name Dynamic Resolution" Support

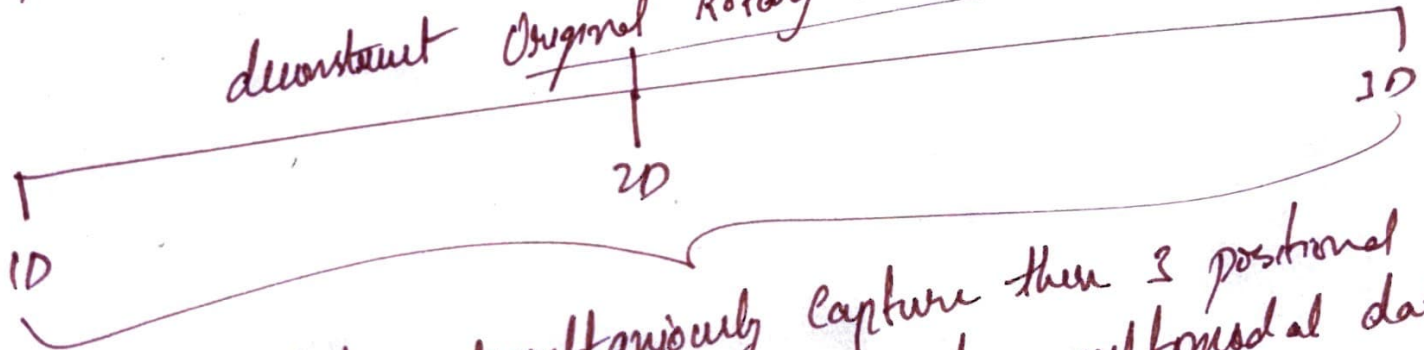


Arbitrary Image Resolution and

Dynamically mapping them into various number of visual tokens.

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

deconstruct Original Rotary Embedding to 3 Categories



helps simultaneously capture three positional information to understand complex multimodal data