Paper Reading(2025.19-present)

Contents

Qwen2-VL

https://arxiv.org/abs/2409.12191

模型结构

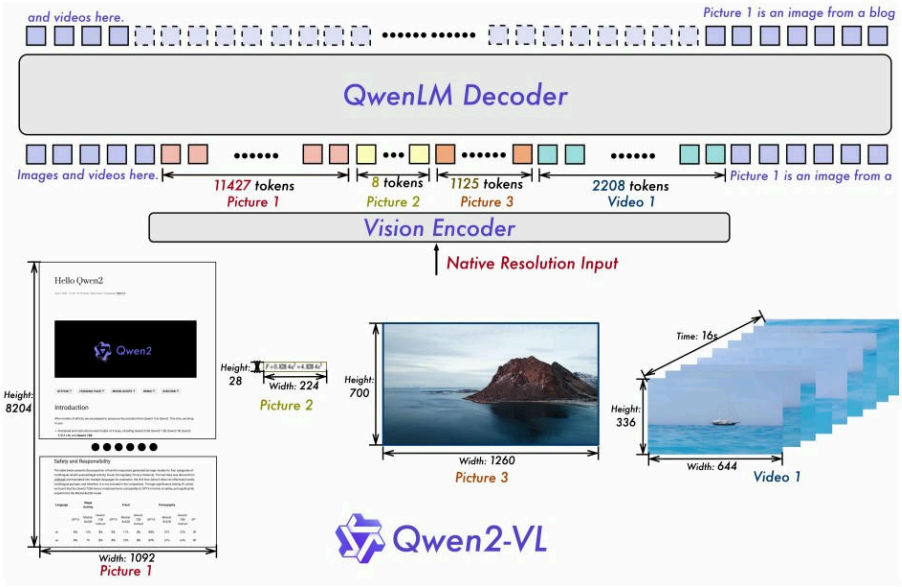

Figure 1: Qwen2-VL structure

- Naive Dynamic Resolution: 可将图片动态转换成若干数量的视觉 tokens，支持任意分辨率。修改 ViT，用 2D-RoPE 代替原本的绝对位置编码嵌入以获取图像的二维信息。在推理阶段，各种分辨率的图像打包成一个序列。为了减少每张图片的视觉 tokens，一个简单的 MLP 层接在 ViT 后，奖2×2的 tokens 压缩成一个 token，用特殊的 tokens`<|vision_start|>`和`<|vision_end|>`放在视觉 tokens 的首尾。由此，一个224×224的图像，经过patch_size=14的 ViT 后得到16×16的 tokens，用 MLP 层压缩得到8×8的 tokens，加上首尾标识符一共64个 tokens。
- Multimodal Rotary Position Embedding (M-RoPE): 相比于 1D-RoPE，将旋转位置编码分成三个模态（将特征维数三等分）：时间(temporal)，高度(height)，宽度(width)。对于文本输入，旋转位置编码保持 1D-RoPE 一样。对于图像，用 M-RoPE 编码，每个视觉 token 时间模态保持一致。对于视频，将其视作一系列的帧，每一帧的时间模态增加，每一帧按图像处理。在多模态场景下，当模型

输入包含多种模态时，每种模态的位置编号初始化方式为：在前一模态的最大位置 ID 基础上加一。

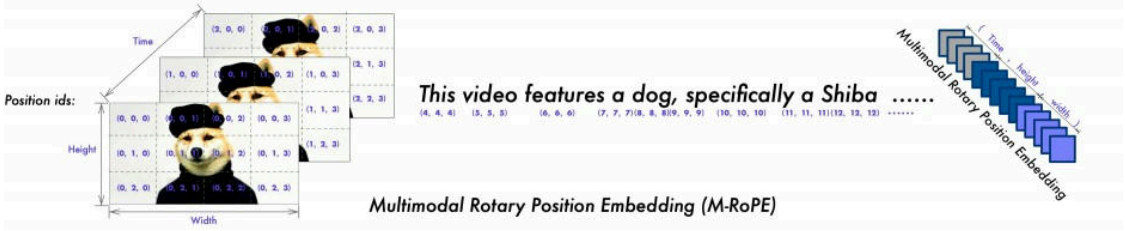

Figure 3: A demonstration of M-RoPE. By decomposing rotary embedding into temporal, height, and width components, M-RoPE can explicitly model the positional information of text, images, and video in LLM.

Figure 2: M-RoPE

- Unified Image and Video Understanding: Qwen2-VL 采用融合图像与视频数据的混合训练方案，确保在图像理解和视频解析两方面都具备专业能力。为了尽可能将视频信息保留完整，对视频每秒采样两帧。Qwen2-VL 集成了深度为 2 的 3D 卷积处理视觉输入，使得 Qwen2-VL 可以处理更多的视频帧而不增加序列长度。为了保持一致性，每张图片视作两个一样的帧（因为同样要用 3D 卷积处理）。为了平衡长视频处理的计算需求与整体训练效率，QWen2-VL 动态调整每帧视频的分辨率，将每个视频的总标记数限制在 16384 个。

训练

采用了 3 阶段的训练方法。

在第一阶段，专注于训练 ViT，通过利用海量图文数据集来增强 LLM 内部的语义理解能力。在第二阶段，解冻所有参数并使用更广泛的数据进行训练，以实现更全面的学习。在最终阶段，锁定 ViT 参数并利用教学数据集对 LLM 进行专项微调。

1. 预训练阶段：数据集包括图像-文本对、OCR、文本-图像交叉的文章、图像问题文本解答、视频对话片段，来源包括清理过的网页、开源数据集、合成数据。预训练阶段集中于学习图像-文本关系、OCR 文本辨识、图像分类工作。这种基础训练有助于使模型形成对视觉-文本核心关联与对齐的深刻理解。

2. 第二个预训练阶段：该阶段引入大量图文混合内容，有助于更细致地理解视觉信息与文本信息之间的交互作用。视觉问答数据集的引入提升了模型处理图像相关查询的能力。此外，多任务数据集的纳入对于开发模型同时处理多样化任务的能力至关重要。这种技能在处理复杂的现实世界数据集时具有极高价值。与此同时，纯文本数据在保持和提升模型语言熟练度方面仍发挥着关键作用。

   在训练过程中，仅对文本标记提供监督。这种接触广泛多样的语言和视觉场景的机制，确保了模型能够深入理解视觉信息与文本信息之间错综复杂的关系，从而为各类多模态任务奠定坚实基础。

3. 微调阶段：采用 ChatML[1]格式构建指令跟随数据。该数据集不仅包含纯文本对话数据，还涵盖多模态会话数据。通过整合多元数据类型，致力于开发一个更通用、更稳健的语言模型，使其不仅能处理传统文本交互，还能胜任复杂的多模态任务。

[1] https://github.com/openai/openai-python/issues/506

Qwen-VL

https://arxiv.org/abs/2308.12966