Qwen-VL

demo:https://tongyi.aliyun.com/qianwen/

github:https://github.com/QwenLM/Qwen-VL

paper:https://arxiv.org/pdf/2308.12966.pdf

特点

• 领先的性能：与相似规模的对手相比，Qwen-VL 在广泛的以视觉为中心的理解基准测试中实现了顶尖的准确性。此外，Qwen-VL 的惊人表现不仅涵盖了传统的基准测试，如图像描述、问答、**定位**等，还包括一些最近引入的对话基准测试。
• **多语言**：与 Qwen-LM 类似，Qwen-VL 是在多语言图像文本数据上进行训练的，其中相当一部分语料库是英文和中文。这样，Qwen-VL 自然支持英文、中文和多语言指令。
• **多图像**：在训练阶段，我们允许任意交错的图像文本数据作为 Qwen-VL 的输入。这个特性使我们的 Qwen-Chat-VL 在给定多个图像时能够比较、理解和分析上下文。
• **精细的视觉理解**：由于我们在训练中使用了更高分辨率的输入大小和精细的语料库，Qwen-VL 表现出高度竞争力的精细视觉理解能力。相比于其他视觉语言通用模型，Qwen-VL 在图像理解方面表现出更好的性能，包括对图像中细微特征的捕捉和理解，以及在各种视觉任务中展现出更高的准确性和全面性。

模型架构

Qwen-VL 的模型架构包括三个组件：大型语言模型(Qwen-7B)、视觉编码器(ViT)和位置感知适配器(single-layer cross-attention)。

Table 1: Details of Qwen-VL model parameters.

Vision Encoder	VL Adapter	LLM	Total
1.9B	0.08B	7.7B	9.6B

数据集细节

Qwen-VL 在图像特征序列的开头和结尾分别添加了两个特殊标记（**<img>和</img>**），表示图像内容的开始和结束。

检测边界框字符串格式："<box>(X_topleft, Y_topleft),(X_bottomright, Y_bottomright)</box>"。

Qwen-VL 在边界框字符串的开头和结尾添加了两个特殊标记（**<box>和</box>**）。此外，为了将边界框与其对应的描述性单词或句子适当地关联起来，Qwen-VL 引入了另一组特殊标记（**<ref>和</ref>**），标记边界框所指的内容。

Data Format of Multi-Task Pre-training

Image Captioning

<img>cc3m/01581435.jpg</img>Generate the caption in English: the beautiful flowers for design.<eos>

Vision Question Answering

<img>VG_100K_2/1.jpg</img> Does the bandage have a different color than the wrist band? Answer: No, both the bandage and the wrist band are white.<eos>

OCR VQA

<img>ocr_vqa/1.jpg</img> What is the title of this book? Answer: Asi Se Dice!, Volume 2: Workbook And Audio Activities (Glencoe Spanish) (Spanish Edition)<eos>

Caption with Grounding

<img>coyo700m/1.jpg</img>Generate the caption in English with grounding: Beautiful shot of <ref>bees</ref><box>(661,612),(833,812)</box><box>(120,555),(265,770) </box> gathering nectars from <ref>an apricot flower</ref><box>(224,13),(399,313) </box><eos>

Referring Grounding

<img>VG_100K_2/3.jpg</img><ref>the ear on a giraffe</ref><box>(176,106),(232,160) </box><eos>

Grounded Captioning

<img>VG_100K_2/4.jpg</img><ref>This</ref><box>(360,542),(476,705)</box> is Yellow cross country ski racing gloves<eos>

OCR

<img>synthdog/1.jpg</img>OCR with grounding: <ref>It is managed</ref> <quad> (568,121), (625,131), (624,182), (567,172)</quad>...<eos>

Data Format of Supervised Fine-tuning

The Dataset Format Example of ChatML

<im_start>user
Picture 1: <img>vg/VG_100K_2/649.jpg</img>What is the sign in the picture?<im_end>
<im_start>assistant
The sign is a road closure with an orange rhombus.<im_end>
<im_start>user
How is the weather in the picture?<im_end>
<im_start>assistant
The shape of the road closure sign is an orange rhombus.<im_end>

训练

Table 8: Training hyperparameters of Qwen-VL

Configuration	Pre-training	Multi-task Pre-training	Supervised Fine-tuning
ViT init.	Open-CLIP-bigG	Qwen-VL 1st-stage	Qwen-VL 2nd-stage
LLM init.	Qwen-7B	Qwen-7B	Qwen-VL 2nd-stage
VL Adapter init.	random	Qwen-VL 1st-stage	Qwen-VL 2nd-stage
Image resolution	224^2	448^2	448^2
ViT sequence length	256	1024	1024
LLM sequence length	512	2048	2048

Qwen-VL 的训练过程包括三个阶段：两个预训练阶段和一个指令微调阶段。

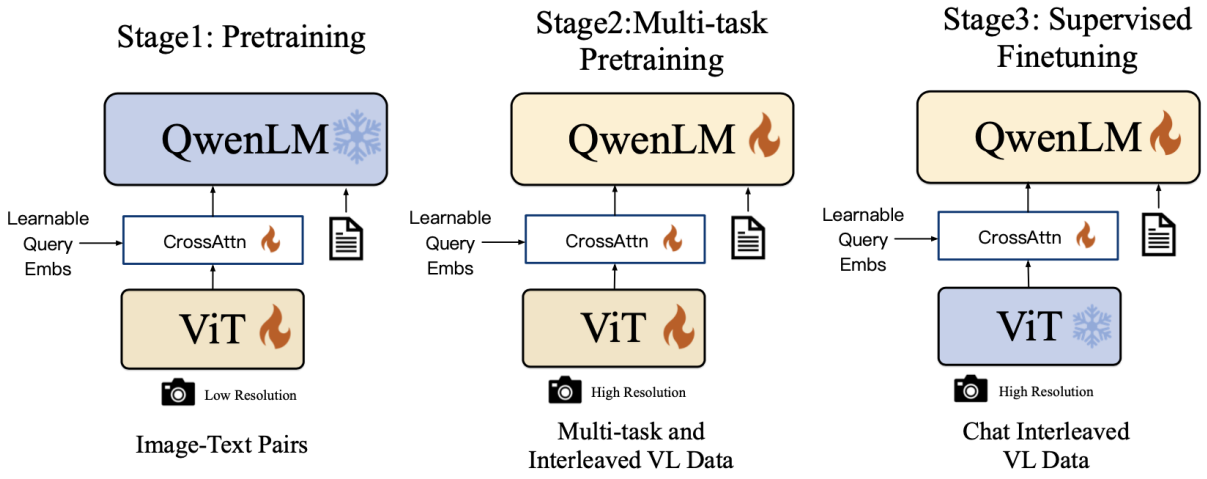

Figure 3: The training pipeline of the Qwen-VL series.

Pre-training

在第一个预训练阶段中，Qwen-VL 主要利用大规模的弱标注网络爬取的图像文本对数据集进行训练。该数据集由几个公开可访问的来源和一些内部数据组成。在清理数据集时，我们努力消除了某些模式。原始数据集包含总共 50 亿个图像文本对，经过清理后，剩下 14 亿个数据，**其中 77.3% 是英文（文本）数据，22.7% 是中文（文本）数据。**

Language	Dataset	Original	Cleaned	Remaining%
English	LAION-en	2B	280M	14%
	LAION-COCO	600M	300M	50%
	DataComp	1.4B	300M	21%
	Coyo	700M	200M	28%
	CC12M	12M	8M	66%
	CC3M	3M	3M	100%
	SBU	1M	0.8M	80%
	COCO Caption	0.6M	0.6M	100%
Chinese	LAION-zh	108M	105M	97%
	In-house Data	220M	220M	100%
	Total	5B	1.4B	28%

Multi-task Pre-training

在第二个预训练阶段中，Qwen-VL 引入了高质量和细粒度的视觉语言注释数据，采用更大的输入分辨率和交错的图像文本数据进行训练。**Qwen-VL 在这个阶段同时训练了 7 个任务，包括文本生成、图像描述、视觉问答(VQA)、视觉定位(Grounding)、参考定位、基于参考的图像描述和文本定位。**

Table 3: Details of Qwen-VL multi-task pre-training data.

Task	# Samples	Dataset
Captioning	19.7M	LAION-en & zh, DataComp, Coyo, CC12M & 3M, SBU, COCO, In-house Data
VQA	3.6M	GQA, VGQA, VQAv2, DVQA, OCR-VQA, DocVQA, TextVQA, ChartQA, AI2D
Grounding[2]	3.5M	GRIT
Ref Grounding	8.7M	GRIT, Visual Genome, RefCOCO, RefCOCO+, RefCOCOg
Grounded Cap.	8.7M	GRIT, Visual Genome, RefCOCO, RefCOCO+, RefCOCOg
OCR	24.8M	SynthDoG-en & zh, Common Crawl pdf & HTML
Pure-text Autoregression	7.8M	In-house Data

Supervised Fine-tuning

在指令微调阶段中，Qwen-VL 使用指令数据集进行微调，以进一步提高模型在特定任务上的性能。这个阶段的目标是让模型更好地理解和执行特定的指令，例如图像描述、问题回答、视觉定位等。

多模态指令微调数据主要来自于通过 LLM 自我训练生成的图像描述数据或对话数据，这些数据通常只涉及单个图像的对话和推理，并且仅限于图像内容理解。为了将定位和多图像理解能力纳入 Qwen-VL 模型，我们通过手动注释、模型生成和策略串联构建了一个额外的对话数据集。我们确认该模型能够有效地将这些能力转移到更广

泛的语言和问题类型上。此外，在训练过程中，我们混合使用多模态和纯文本对话数据，以确保模型在对话能力方面的普适性。**指令微调数据总共有 350k 条**。

未来工作

未来，作者致力于在以下几个关键维度进一步增强 Qwen-VL 的能力：

- 将 Qwen-VL 与更多的模态集成，例如语音和视频。
- 通过增加模型大小、训练数据和更高的分辨率来扩展 Qwen-VL，使其能够处理更复杂和复杂的多模态数据关系。
- 扩展 Qwen-VL 在多模态生成方面的能力，特别是在生成高保真度图像和流畅语音方面。