11.3 试用DeepseekOCR 副本

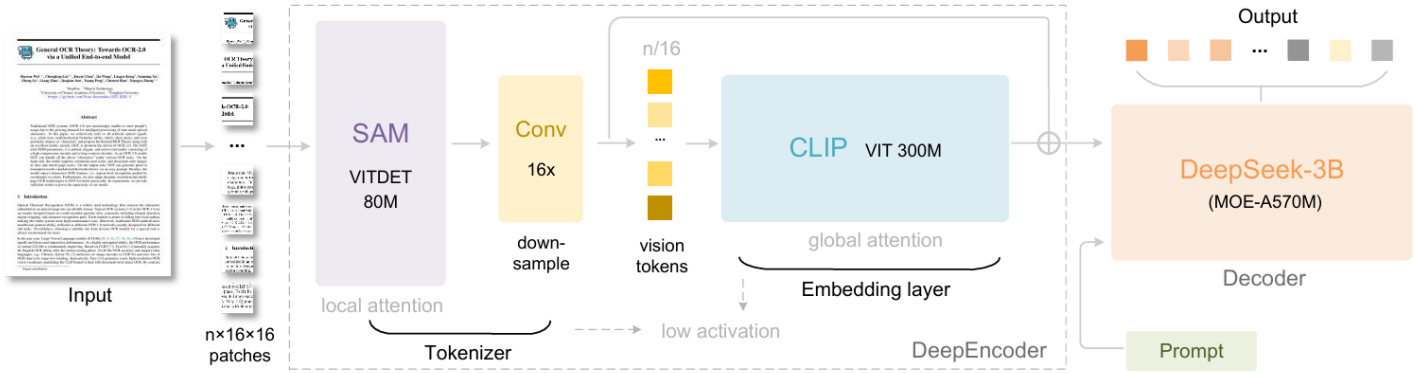
《DeepSeek-OCR: Contexts Optical Compression》

核心目标与背景

当前的大型语言模型（LLMs）在处理长文本内容时，面临着巨大的计算挑战。作者提出了一种潜在的解决方案：利用视觉模态作为文本信息的高效压缩媒介。通过将文档文本转化为图像，可以用比等效数字文本少得多的视觉 Token 来表示丰富的信息，从而实现更高的压缩率。**DeepSeek-OCR** 是一个初步的概念验证模型，旨在高效地实现视觉-文本压缩。

模型架构

DeepSeek-OCR 采用统一的端到端视觉-语言模型（VLM）架构，包含两个主要组件：



DeepEncoder

提取图像特征、Token 化并压缩视觉表示。它是 DeepSeek-OCR 的核心引擎。旨在在高分辨率输入下保持低激活内存，同时实现高压缩比和可控的视觉 Token 数量。由一个基于 **SAM-base**（主要用于感知、以窗口注意力为主）和一个基于 **CLIP-large**（主要用于视觉知识、以密集全局注意力为主）的组件串联而成。在两个组件之间，使用一个 **16 倍的 Token 压缩器**（由 2 层卷积模块实现）进行 Token 数量的降采样，从而使整体激活内存可控。

DeepSeek3B-MoE (解码器)

根据图像 Token 和提示词生成所需的文本结果。采用 **DeepSeek-3B-MoE** 架构，在推理时激活6个路由专家和 2 个共享专家，激活参数约为 **570M**。这种设计在保持 3B 模型表达能力的同时，实现了 500M 小模型的推理效率 。

关键实验结果

1.在 Fox 基准测试中，DeepSeek-OCR 展示了卓越的压缩-解压缩能力。

- **10 倍压缩比以下：** 当文本 Token 数量在视觉 Token 数量的 **10 倍以内**时（即压缩比 < 10x），模型可以实现 **97%** 的 OCR 解码精度。
- **20 倍压缩比：** OCR 准确率仍能保持在约 **60%**。

Text Tokens	Vision Tokens =64		Vision Tokens=100		
	Precision	Compression	Precision	Compression	Pages
600-700	96.5%	10.5×	98.5%	6.7×	7
700-800	93.8%	11.8×	97.3%	7.5×	28
800-900	83.8%	13.2×	96.8%	8.5×	28
900-1000	85.9%	15.1×	96.8%	9.7×	14
1000-1100	79.3%	16.5×	91.5%	10.6×	11
1100-1200	76.4%	17.7×	89.8%	11.3×	8
1200-1300	59.1%	19.7×	87.1%	12.6×	4

核心功能

OCR纯文字提取：

支持对任意图像进行自由式文字识别（Free OCR），快速提取图片中的全部文本信息，不依赖版面结构，适合截图、票据、合同片段等轻量场景的快速文本获取。

保留版面格式的OCR提取：

模型可自动识别并重建文档中的排版结构，包括段落、标题、页眉页脚、列表与多栏布局，实现“结构化文字输出”。此功能可直接将扫描文档还原为可编辑的排版文本，方便二次编辑与归档。

图表 & 表格解析：

DeepSeek-OCR 不仅识别文本，还能解析图像中的结构化信息，如表格、流程图、建筑平面图等，自动识别单元格边界、字段对齐关系及数据对应结构，支持生成可机读的表格或文本描述。

图片信息描述：

借助其多模态理解能力，模型能够对整张图片进行语义级分析与详细描述，生成自然语言总结，适用于视觉报告生成、科研论文图像理解以及复杂视觉场景说明。

指定元素位置锁定：

支持通过“视觉定位”（Grounding）功能，在图像中准确定位特定目标元素。例如，输入“Locate signature in the image”，模型即可返回签名区域的坐标，实现基于语义的图像检索与目标检测。

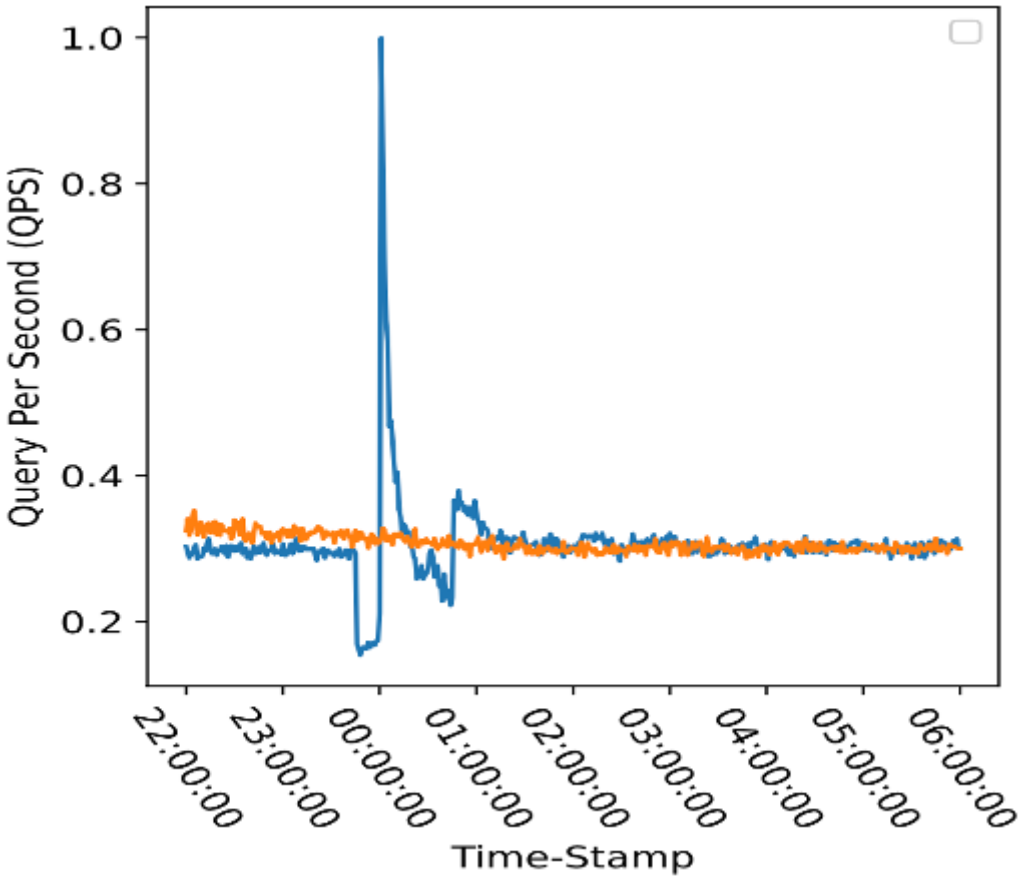
Markdown文档转化：

可将完整的文档图像直接转换为结构化 Markdown 文本，自动识别标题层级、段落结构、表格与列表格式，是实现文档数字化、知识库构建和多模态RAG场景的重要基础模块。

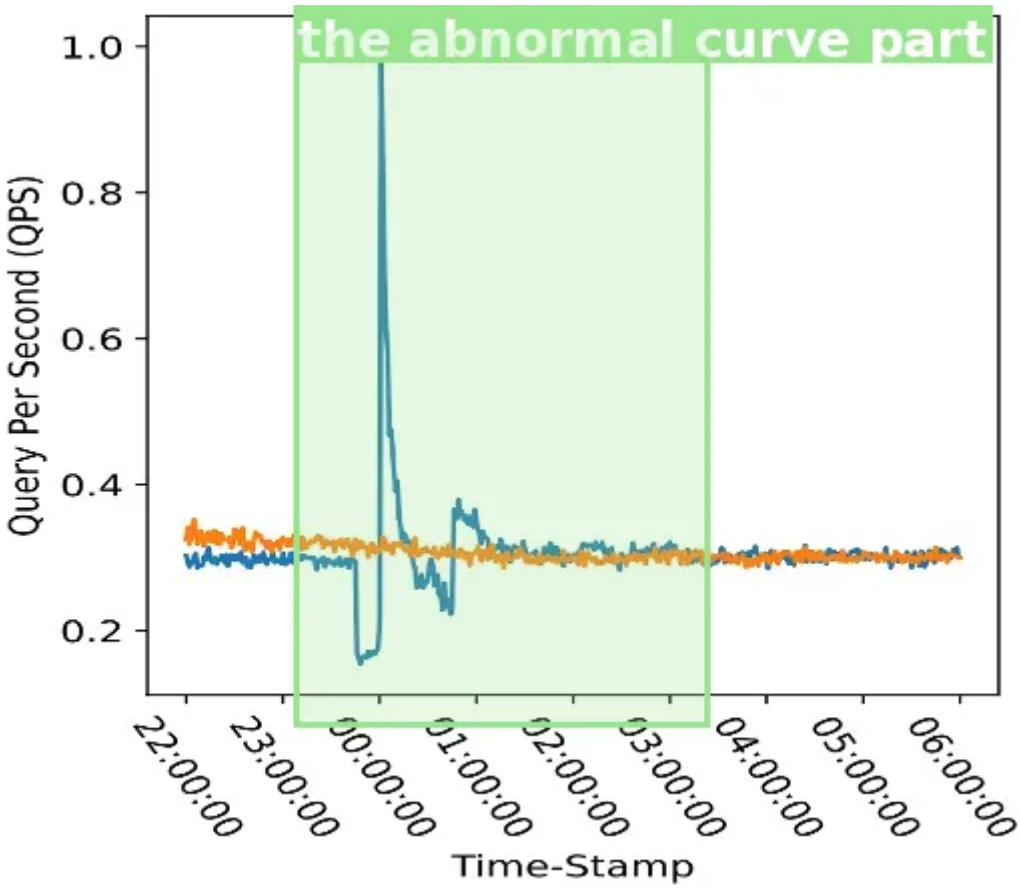
如何用在微服务异常检测任务中

输入

Prompt: Locate the abnormal curve part.



输出



注：遵循 DeepSeek-VL2，我们为 caption、detection 和 grounding 等任务生成相关数据。需要注意的是，DeepSeek-OCR 并不是通用的视觉语言模型（VLM），这部分数据仅占总数据的 20%。我们引入这种类型的数据主要是为了保留通用的视觉接口，以便对我们的模型和通用视觉任务感兴趣的研究人员在未来能够方便地推进他们的工作。