接口参考

配置示例

模型	配置	状态	操作	协议

multilingual-e5-large	{
	"parameters": {
	"type": "dataelem.pymodel.huggingface_model",
	"pymodel_type": "embedding.ME5Embedding",
	"gpu_memory": "3",
	"instance_groups": "device=gpu;gpus=7|8"
	}
	}
		ready/unready	on/off/reload	MIT

bge-large-zh	{
	"parameters": {
	"type": "dataelem.pymodel.huggingface_model",
	"pymodel_type": "embedding.BGEZhEmbedding",
	"gpu_memory": "3",
	"instance_groups": "device=gpu;gpus=7"
	}
	}
		ready/unready		MIT

gte-large	{
	"parameters": {
	"type": "dataelem.pymodel.huggingface_model",
	"pymodel_type": "embedding.GTEEmbedding",
	"gpu_memory": "3",
	"instance_groups": "device=gpu;gpus=7"
	}
	}
		ready/unready		MIT

	{			

Baichuan-13B-Chat	"parameters": {
	"type": "dataelem.pymodel.huggingface_model",
	"pymodel_type": "llm.BaichuanChat",
	"pymodel_params": "{\"max_tokens\": 4096}",
	"gpu_memory": "30",
	"instance_groups": "device=gpu;gpus=7,8"
	}
	}
				Authorized

chatglm2-6b
chatglm2-6b-32k
	{
	"parameters": {
	"type": "dataelem.pymodel.huggingface_model",
	"pymodel_type": "llm.ChatGLM2",
	"gpu_memory": "15",
	"instance_groups": "device=gpu;gpus=7"
	}
	}
				Authorized

Llama-2-13b-chat-hf	{
	"parameters": {
	"type": "dataelem.pymodel.huggingface_model",
	"pymodel_type": "llm.Llama2Chat",
	"gpu_memory": "30",
	"instance_groups": "device=gpu;gpus=7,8"
	}
	}
				Authorized

Llama-2-7b-chat-hf	{
	"parameters": {
	"type": "dataelem.pymodel.huggingface_model",
	"pymodel_type": "llm.Llama2Chat",
	"gpu_memory": "15",
	"instance_groups": "device=gpu;gpus=7"
	}
	}
				Authorized

Qwen-7B-Chat	{
	"parameters": {
	"type": "dataelem.pymodel.huggingface_model",
	"pymodel_type": "llm.QwenChat",
				Authorized

	"gpu_memory": "15", "instance_groups": "device=gpu;gpus=7" } }			

visualglm-6b	{
	"parameters": {
	"type":
	"dataelem.pymodel.huggingface_model",
	"pymodel_type": "mmu.VisualGLM",
	"gpu_memory": "16",
	"instance_groups": "device=gpu;gpus=7"
	}
	}
				The VisualGLM-6B License

elem_layout_v1	{
	"parameters": {
	"type": "dataelem.pymodel.elem_model",
	"pymodel_type": "layout.LayoutMrcnn",
	"gpu_memory": "4",
	"instance_groups": "device=gpu;gpus=6"
	}
	}
				DataElem, Inc.

elem_table_detect_v1	{
	"parameters": {
	"type": "dataelem.pymodel.elem_model",
	"pymodel_type": "table.MrcnnTableDetect",
	"gpu_memory": "4",
	"instance_groups": "device=gpu;gpus=6"
	}
	}
				DataElem, Inc.

elem_table_cell_detect_v1	{
	"parameters": {
	"type": "dataelem.pymodel.elem_model",
	"pymodel_type": "table.TableCellApp",
	"gpu_memory": "4",
	"instance_groups": "device=gpu;gpus=6"
	}
	}
				DataElem, Inc.

elem_table_rowcol_detect_v1	{
	"parameters": {
	"type": "dataelem.pymodel.elem_model",
	"pymodel_type": "table.TableRowColApp",
				DataElem, Inc.

	"gpu_memory": "4", "instance_groups": "device=gpu;gpus=6" } }			

模型配置参数说明

格式： json {"parameters": {字段名: 字段值}}，字段值都是string类型，支持的字段和字段值说明如下：

字段名	含义	默认值、是否必填
type	str，表示模型所属的类型，两类可选 [dataelem.pymodel.huggingface_model, dataelem.pymodel.elem_model] 前者表示huggingface的模型格式，对应huggingface的模型仓库要求格式 后者表示DataElemu定义的模型格式，规范如下： tensorflow模型： 模型文件夹/{model.graphdef, model_def.json} model.graphdef是freeze格式的模型 pytorch模型：模型文件夹/{model.pt, model_def.json}	无，必填
pymodel_type	str，表示具体的模型类，支持如下类型 • embedding.ME5Embedding • embedding.BGEZhEmbedding • embedding.GTEEmbedding • llm.BaichuanChat • llm.ChatGLM2 • llm.Llama2Chat • llm.QwenChat • mmu.VisualGLM • layout.LayoutMrcnn • table.MrcnnTableDetect • table.TableCellApp • table.TableRowColApp • ocr.PPOCRCollectionV3 [V4] (待支持) • ocr.ElemOCRColletionV1（待支持） • nlp.ElmModelV1 （待支持）	无，必填

pymodel_params	str, 表示模型的初始化参数，要求是一个符合json格式的字符串 例如"{\"max_tokens\": 4096}"	"{}"，选填
gpu_memory	str，表示模型需要的gpu显存总大小，单位为GB	无，必填
instance_groups	str，表示gpu的分配策略 格式要求：device=gpu;gpus=id1[,id2]+[|id3[,id4]+] 举例说明： - device=gpu;gpus=0 表示在0卡申请gpu_memory大小的显存 - device=gpu;gpus=0,1表示在0,1卡每张卡申请gpu_memory/2大小的显存 - device=gpu;gpus=0,1,2表示在0,1,2卡每张卡申请gpu_memory/3大小的显存 - device=gpu;gpus=0|1|2 表示在0,1,2卡上各自启动1个模型推理实例，并且每张卡上申请gpu_memory大小的显存 - device=gpu;gpus=0,1|1,2 表示启动2个模型推理实例，第一个实例需要使用0，1卡，第二个实例需要1，2卡	无，必填
precision	str，表示推理的精度，默认是fp16，支持fp16,fp32 是否支持int4,int8依赖特定模型，一般不推荐使用	“fp32”，选填