torchserve::ModelInstance

+ uint16_t intanceId

+ virtual Status createInstance (const Model &model, int gpuId)=0
+ void predict(std::shared _ptr< BatchJob > &job)=0
+ virtual void Warmup()=0
+ uint16_t GetInstanceId()

torchserve::ModelInstanceOMP

- uint32_t port

+ ModelInstanceOMP(uint16 _t instanceId, uint32_t port)
+ Status createInstance (const Model &model, int gpuId)
+ void predict(std::shared _ptr< BatchJob > &job)
+ void warmup()
+ uint32_t getRemotePort()
- Status connectBackend (uint32_t port)