- - Inference request - >
HTTP POST

Temporary data storage

HTTP interface

< - unpack data - >

Pre-processing chain

<- - data - ->

< - do inference - >

Forward Pass

<- - output - ->

< - post process - >

Post-processing chain

Temporary data storage

<- inference response - -

Temporary data delete