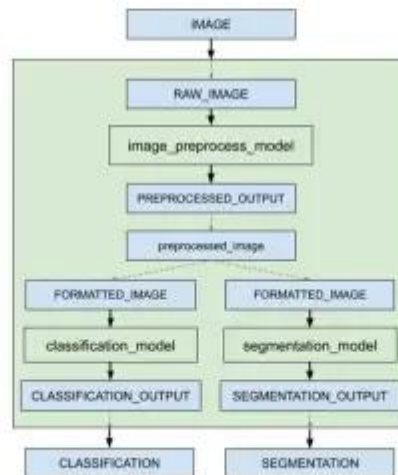


组合模型配置 (ensemble model)

例子 1

CONFIGURE AN ENSEMBLE MODEL

An Example



图片-预处理模型-分别进入不同的模型-输出不同的结果。

CONFIGURE AN ENSEMBLE MODEL

First Glance

```
name: "ensemble_model"
platform: "ensemble"
max_batch_size: 1
input [
  {
    name: "IMAGE"
    data_type: TYPE_STRING
    dims: [ 1 ]
  }
]
output [
  {
    name: "CLASSIFICATION"
    data_type: TYPE_FP32
    dims: [ 1000 ]
  },
  {
    name: "SEGMENTATION"
    data_type: TYPE_FP32
    dims: [ 3, 224, 224 ]
  }
]
```

```
ensemble_scheduling {
  step [
    {
      model_name: "image_preprocess_model"
      model_version: -1
      input_map {
        key: "RAW_IMAGE"
        value: "IMAGE"
      }
      output_map {
        key: "PREPROCESSED_OUTPUT"
        value: "preprocessed_image"
      }
    },
    {
      model_name: "classification_model"
      model_version: -1
      input_map {
        key: "FORMATTED_IMAGE"
        value: "preprocessed_image"
      }
      output_map {
        key: "CLASSIFICATION_OUTPUT"
        value: "CLASSIFICATION"
      }
    },
    {
      model_name: "segmentation_model"
      model_version: -1
      input_map {
        key: "FORMATTED_IMAGE"
        value: "preprocessed_image"
      }
      output_map {
        key: "SEGMENTATION_OUTPUT"
        value: "SEGMENTATION"
      }
    }
  ]
}
```

定义模型的输入输出，然后在 `ensemble_scheduling` 中定义不同的步骤，其中 `step` 中的 `key` 是本身的 `input/output tensor` 的名字；`value` 是 `ensemble model` 中的 `Tensor` 名字。

配置写完后，在 `ensemble_model` 的目录只能够新建一个版本目录，里面为空，然后放 `config` 文件。

注意事项：

CONFIGURE AN ENSEMBLE MODEL

Notices

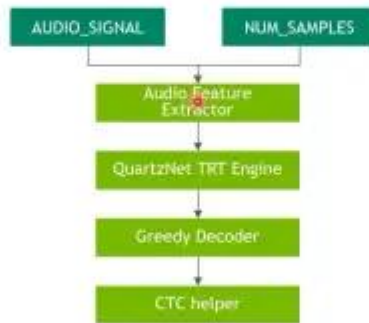
- If one of the models is stateful model, then the inference request should contain the information mentioned in [Stateful Models](#)
- The models composing the ensemble have their own scheduler
- If models in ensemble are all framework backends, data transmission between them does not have to go through CPU memory

- 如果组合里有一个是 `stateful` 模型，那么整个 `pipeline` 都成为 `stateful` 模型，推理请求需要符合 `stateful model` 的规则。
- 每个子模块有各自的调度器。
- 如果每个子模块都是 `framework backend`，则传输使用 `GPU` 进行，否则可能通过 `cpu` 内存。

例子 2

CONFIGURE AN ENSEMBLE MODEL

Example 2

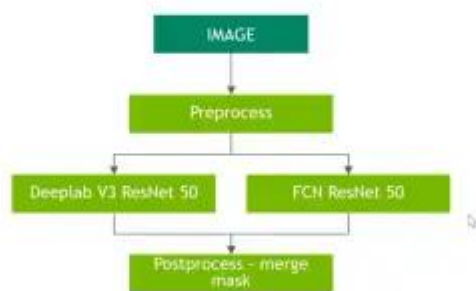


语音识别流程。

例子 3

CONFIGURE AN ENSEMBLE MODEL

Example 3



预处理-分支 1 分割模型-分支 2fcn 分割-合并拼接