web demo 后端配置

在`web_app.py`顶部有一个 BACKENDS 字典，包含四个后端配置。请在`url`中输入对应的 http 链接

- 目前本地模型默认使用的是跟其他平台一样走http的配置，暂时没讲其显示在web的可选项中，如果需要修改的话我后面改一下
- HTTP 方法：`POST`
- 请求体

```
{ "image": "<base64 编码的图片字节>" }
```

- 响应体 至少要有

```
{ "prediction": "cat" }
```

对于测试文件的读取我目前假设csv的保存格式如下，如果需要调整请跟我说

- 列名：`backend, avg_latency_ms, avg_ips, cost_per_1m`
  - backend：lambda / ec2_x86 / ec2_arm 之一
  - avg_latency_ms：平均延迟
  - avg_ips：平均吞吐
  - cost_per_1m：每 1M 推理成本（美元）

web demo运行

1. web使用flask实现，首先请确保安装依赖

```
 pip install flask requests
```

2. 进入web目录使用以下命令启动脚本

```
python web_app.py
```

3. 浏览器访问`http://127.0.0.1:8000`