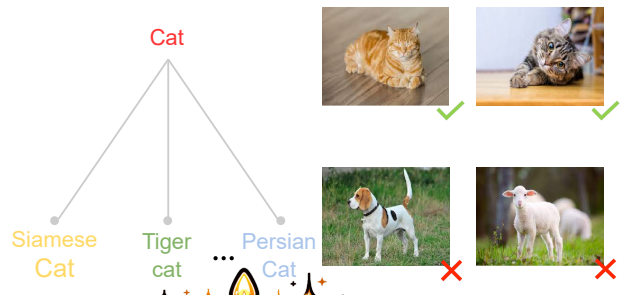
Request

We require a MobileNet-based model which has been specifically trained to differentiate between breeds of domestic cats including the Siamese cat, Persian cat, and the Tiger cat. The model should maintain an accuracy of over 98%. Please ensure that the total number of model parameters does not exceed 20M. Futhuremore, the model should be deployed on a GPU using the ONNX Runtime inference engine and process at least 100 frames per second.

Stage #1: Request Understanding

Data:
  Targets: Siamese cat, Persian cat,Tiger cat
Model:
  Task: Image Classification. **Name**: MobileNet
  Params: 20M **FLOPs**: none. **Speed**: 100fps
Deploy:
  Device: GPU **Inference engine:** onnxruntime
  ...

Stage #2: Data Selection



Stage #3: Model Selection

Name: mobile-v3-large
Task: classification
Params(M): 5.48
FLOPs(G): 0.23
Speed(fps): 174.42

Name: swin-base
Task: classification
Params(M): 49.61
FLOPs(G): 8.52
Speed(fps): 37.72

Stage #4: Model Training with HPO

HPO-LLaMA

Optimizer: Adam
Initial Lr: 1e-4
WeightDecay: 1e-4
Batchsize: 4
…

Model Training
MMPreTrain MMPose
MMDetection MMSeg

Stage #5: Model Deployment

MMDeploy

OpenVINO ONNX RUNTIME TensorRT C++
CANN Qualcomm Rockchip 瀚博半导体 POWERED BY VASTSTREAM

Response

Request Understand: The requirements parsing is completed, and the parsing results are as follows ...

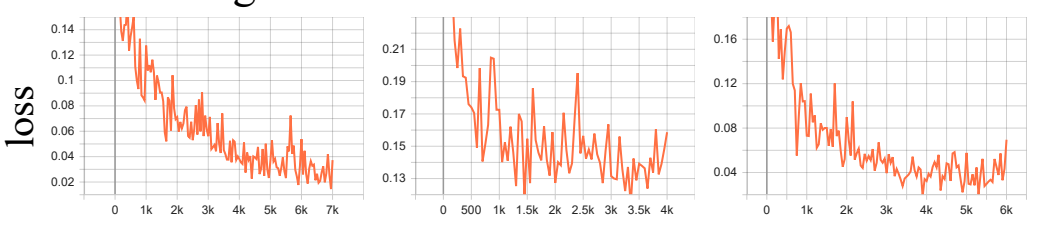
Data Selection: According to your request, I collect 3900 related images from the ImageNet dataset and construct a training dataset. The number of images in each category in the training dataset is as follows:

Siamese Cat	Persian Cat	Tiger cat
1300	1300	1300

Model Selection: Based on your request, I find the following model that meets your requirements:

Model	Params(M)	Flops(G)	Inf time (fps)
mobilenet-v3-large	5.48	0.23	174.42

Model Training and Hyperparameter Optimization: The model is training ...



After three rounds of hyperparameter optimization, the accuracy of the model on the test set reaches 98.01%, which meets the requirements.

Model Deploy: The model was successfully converted to ONNX Runtime format, please download it in the link: http://127.0.0.1:10045/model.zip