CENG5030 Lab 04

Introduction to Distiller

1 Intro to Distiller

Steps

- Install Distiller, create and activate a virtual environment `env`
- Make sure that you're at the root directory of `Distiller`
- Define your own model:
  - Add your model into `./distiller/models/cifar10`, an example: `./Lab04-code/ceng5030_cifar.py`
  - Register your model at the end of `./distiller/models/cifar10/__init__.py`: `from .ceng5030_cifar import *`
- Step into `./examples/classifier_compression`
- Train your model from scratch: run "`python3 compress_classifier.py --arch ceng5030_cifar ../../../data.cifar10 -p 30 -j=1 --lr=0.01 --epochs 2`" (the dataset will be downloaded automatically)
- Check your log file to locate the checkpoint file
- Evaluate your model: run "`python3 compress_classifier.py --arch ceng5030_cifar ../../../data.cifar10 -p=50 --resume-from=XXXX --evaluate`", here `XXX` is your checkpoint file
- Define your pruning algorithm:
  - Add your algorithm into `./distiller/pruning`, an example: `./Lab04-code/ceng5030_pruner.py`
  - Register your algorithm in `./distiller/pruning/__init__.py`: `from .ceng5030_pruner import CENG5030Pruner`
- Prune your model:
  - Load the stored checkpoint file to check the weights, an example code is in `./Lab04-code/checkname.py`; or you can get them by evaluating the model
  - Add your own pruning configuration file into `./examples/sensitivity-pruning`
  - Run "`python3 compress_classifier.py --arch=ceng5030_cifar ../../../data.cifar10 -p=50 --lr=0.1 --epochs=2 --batch=128 --compress =../sensitivity-pruning/ceng5030_cifar.schedule_sensitivity.yaml --gpu=0 -j=1 --deterministic`" to prune the model

2 Sample Code

- check the model weights: `./Lab04-code/checkname.py`
- Define your model:
  - model file: `./Lab04-code/ceng5030_cifar.py`
  - register your model: "`from .ceng5030_cifar import *`"
- Define your pruning algorithm:
  - algorithm file `./Lab04-code/ceng5030_pruner.py`
  - register your algorithm: "`from .ceng5030_pruner import CENG5030Pruner`"
- pruning schedule file: `./Lab04-code/ceng5030_cifar.schedule_sensitivity.yaml`
- scripts: `./Lab04-code/lab04.sh`

3 Assignment

Q1 Prune the squeezenet1_1 (provided by Distiller, python ./compress_classifier.py -h)

- Dataset: ImageNet (Tiny ImageNet)
- Use group dependency type `Leader` with at least one dependency group
- Use two `group types`: `Channels` and `Filters`.
- Use two `pruning algorithms`: `Automated Gradual Pruner (AGP)` and `sensitivity pruning`
- You do not need to cover all of these types or algorithms in a single YAML file, in other words, you can implement these algorithms individually in several configuration YAML files.

Q2 Learn to quantize MobileNet (provided by Distiller)

- Dataset: ImageNet (Tiny ImageNet)
- Quantize the model via 8-bit post-training linear quantizer.

Useful Materials:

- You can use Tiny-ImageNet
- PyTorch models
- Netron
- named_parameters()
- A bug occured when resuming from the store model