BASIC LAB

# Data Science
# HW 2: Model Compression
# (updated time: 03/14)

TA: 曾偉倫

Email: wltseng.ee06@nycu.edu.tw

2023 Spring Data Science

NYCU

# Outline

- Introduction – Model Compression
- Problem Description
- Kaggle Competition
- Grading Policy
- Report & Demo
- E3 Submission

# Introduction

- Model Compression
  - Knowledge Distillation
  - Pruning
  - Model Architecture Design

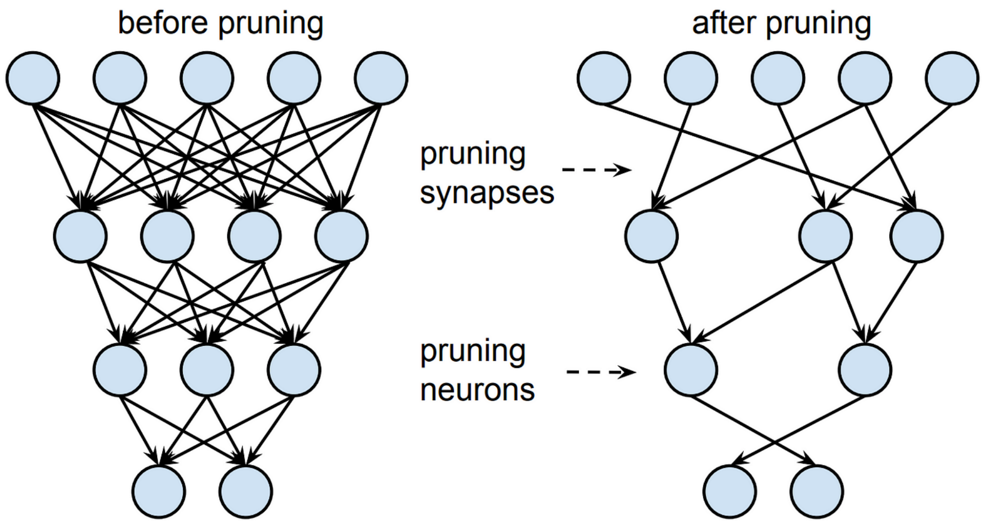

Figure. Pruning (ref)

# Problem Description

- Dataset: Fashion MNIST
- Input: Well-trained ResNet-50
- Output: compressed model
- Constrain:
  - number of parameter $\leq$ 100,000
  - accuracy $\geq$ 0.8389
  - DO NOT USE ANY TEST DATA, EXTERNAL DATA

# Clarification (3/14)

BASIC LAB

## Released Resnet-50 model

- It is not permitted to train your own teacher model.
- You can only use the released resnet-50.pth as the pre-trained weight for your teacher model to compress.

## Example Case:

1. You can use the released resnet-50.pth to distill it into a smaller model and then compress it (**V**).
2. It is not allowed to train your own teacher model and distill/prune it later (**X**).
3. It is not allowed to fine-tune the released resnet-50.pth to improve the teacher model and then compress it (**X**).

# 情境舉例

情境1. : 我自己訓練一個fashion MNIST的model，再進行compression。

Ans: 不行。這次題目是要比較壓縮的方法，如果壓縮的基準點不一樣，就失去公平性。

情境2.：我想把助教release的model進行finetune後再進行壓縮。

Ans: 不行，這樣跟其他同學比較基準點就不一樣。助教在demo重現結果時，會檢查是否有額外再finetune原始的resnet-50。

# 情境舉例

情境3.：我想修改model architecture來進行壓縮，使用depthwise-conv 之類的方式來壓縮模型。

Ans: 可以，但是一樣要使用release的model來當作pretrained weight。(將原來卷積層的權重轉移至新的depthwise卷積層)

情境4.：我想使用torch.nn.prune以外的套件來實作compressio可以嗎?
Ans: 可以，請在report中說明來源、使用方法。

# 情境舉例

情境5.：我使用助教給的resnet-50模型，distill到自己設計的小model後再進行壓縮。

Ans: 可以。符合題目的要求，使用release的model來進行壓縮。

情境6.：student model可以使用imagenet的pretrained weigh來initializ嗎?
Ans: 可以，student的部分沒有規定，讓同學自由發揮。

# Check Number of parameter (03/14)

- Please use “torchinfo” python package if you use “nn.torch.prune”
  - Feature: support to check “nn.torch.prune” results
  - Install & check your model paprameter:
    1. Commandline to install torchinfo
       ```
       pip install torchinfo
       ```
    2. Import torchinfo:
       ```
       import torchinfo as summary
       ```
    3. Use torchinfo summary to see result:
       ```
       summary(model)
       ```

# Grading Policy

Model Compression (total: 100%)

- Kaggle Competition (75%)
- Report (20%)
- Demo (5%)

# Grading Policy

Model Compression (total: 100%)

- Kaggle Competition (45%+30%)
  - Constrain: num_parameter <= 100,000
  - 45%: accuracy ≥ baseline benchmark (update to Kaggle soon)
  - 30%: private leaderboard ranking

# Kaggle Competition

- Invitation Link: https://www.kaggle.com/t/ee36089663ee48cba68845dd1b791fba

- A maximum of 5 submissions per day is allowed on Kaggle.
- Timeline:
  - 3/07 12:00 Competition Start
  - 3/20 23:59 Competition Finished

# Grading Policy

- Report (20%)
  - torchsummay/torchinfo output (5%)
  - Brief Explanation of Compression Methods (15%)
    - Name, student_ID
    - Methods you used
    - Reference
    - $\leq$ 200 words
- Demo (5%)
  - TA will execute your code and reproduce the results.

# Special Rules

1. Plagiarism is prohibited.
2. Sharing of code or submission files is prohibited.
3. A maximum of 5 submissions per day is allowed on Kaggle.
   Please do not use any methods to bypass this limit.
4. Using testing data or external data is prohibited. TA will check the dataloader.
5. Using pre-trained models created by others as the final result is prohibited.
   Please train your own model.
6. Using other models for compression is prohibited. Please use the trained model provided in the assignment release.

Violation of any of the above rules will result in a score of 0 for this assignment.

# Demo Platform

- OS: Ubuntu Server 20.04
- CPU: AMD Ryzen Threadripper  (will set num_worker=8)
- GPU: RTX 3080 (8GB) *1
- Python 3.8.10
- CUDA: 11.07
- Framework: PyTorch 1.13.1

# E3 Submission

- Two File:
  1. <pdf file> hw2_report_[student_ID].pdf
     - Example: “hw2_report_311000123.pdf”
  2. <zip file> hw2_[student_ID].zip
     - Example: “hw2_311000123.zip”
     - Please make sure your submission contains the following items:
       1) All the code you used for training and testing
       2) The final weights used for testing
       3) A README file explaining how to execute your code (e.g., in txt or md format)

# Demo Platform (計中平台的同學)

- OS: Ubuntu Server 20.04
- CPU: AMD Ryzen Threadripper  (will set num_worker=8)
- GPU: RTX 3080 (8GB) *1
- Python 3.8.10
- CUDA: ~~11.07~~ 10.1
- Framework: ~~PyTorch 1.13.1~~  (PyTorch 1.8.1)

# 補充說明

BASIC LAB

Q: 有同學反應，使用計中GPU server會遇到CUDA版本過舊，無法使用pytorch 1.13.1。

A: 經過助教與同學確認後，請同學在計中server環境，使用virtualenv後安裝。

並在report註明，使用的pytorch的版本是1.8.1。

助教基本上會以1.13.1的版本為主，1.8.1的版本目前測試後resnet-50的accuracy落差小於0.01%。

``` # 安裝指令：

pip install torch==1.8.1+cu101 torchvision==0.9.1+cu101 torchaudio==0.8.1 -f https://download.pytorch.org/whl/torch_stable.html

```

NYCU

# 使用server的推薦工具

BASIC LAB

- tmux : 04 - Tmux - 終端機管理工具 - iT 邦幫忙::一起幫忙解決難題，拯救 IT 人的一天 (ithome.com.tw)
- Filezilla: [無料才是王道] FTP檔案傳輸 - Filezilla - iT 邦幫忙::一起幫忙解決難題，拯救 IT 人的一天 (ithome.com.tw)
- vscode: [教學] 使用 Visual Studio Code 透過 SSH 進行遠端程式開發 | 辛比誌 (xenby.com)
- Putty: PuTTY v0.78 最多人用的 Telnet, SSH…伺服器連線工具（+中文版） – 重灌狂人 (briian.com)
- MobaXterm: Linux環境搭建 | 全能終端神器——MobaXterm | IT人 (iter01.com)
- Notepad++ with NppFTP: How to Connect to Notepad++ FTP: A Step By Step Guide (hostinger.com)

NYCU