BASIC LAB

# Data Science HW #4
# Model Compression for LLM

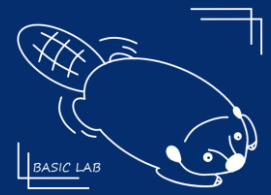
TA: 曾偉倫

Email: wltseng.ee06@nycu.edu.tw

NYCU

# Table of Contents


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


BASIC LAB

NYCU

# Introduction

BASIC LAB

## Model Compression Goals

- Smaller Size
  - Compress Mobile App Size
- Accuracy
  - no loss of accuracy improved accuracy
- Speedup
  - make inference faster

# Tradeoff for Network Compressions

**Model Performance**


**Compression Rate**

https://moptt.tw/p/Beauty.M.1534834822.A.1C5

# Summarization: two main strategies

BASIC LAB

## Extractive summarization

*Select parts* *(typically sentences) of the original text to form a summary.*


- Easier
- Restrictive (no paraphrasing)

## Abstractive summarization

*Generate new text* *using natural language generation techniques.*

- More difficult
- More flexible (more human)



NYCU

# Text Summarization

BASIC LAB

- Example
  - Input :

"summarize: The Inflation Reduction Act lowers prescription drug costs, health care costs, and energy costs. It's the most aggressive action on tackling the climate crisis in American history, which will lift up American workers and create good-paying, union jobs across the country. It'll lower the deficit and ask the ultra-wealthy and corporations to pay their fair share. And no one making under $400,000 per year will pay a penny more in taxes.“

  - Label:

"the Inflation Reduction Act lowers prescription drug costs, health care costs, and energy costs. It's the most aggressive action on tackling the climate crisis in American history, which will lift up American workers and create good-paying, union jobs across the country."

# Network Pruning

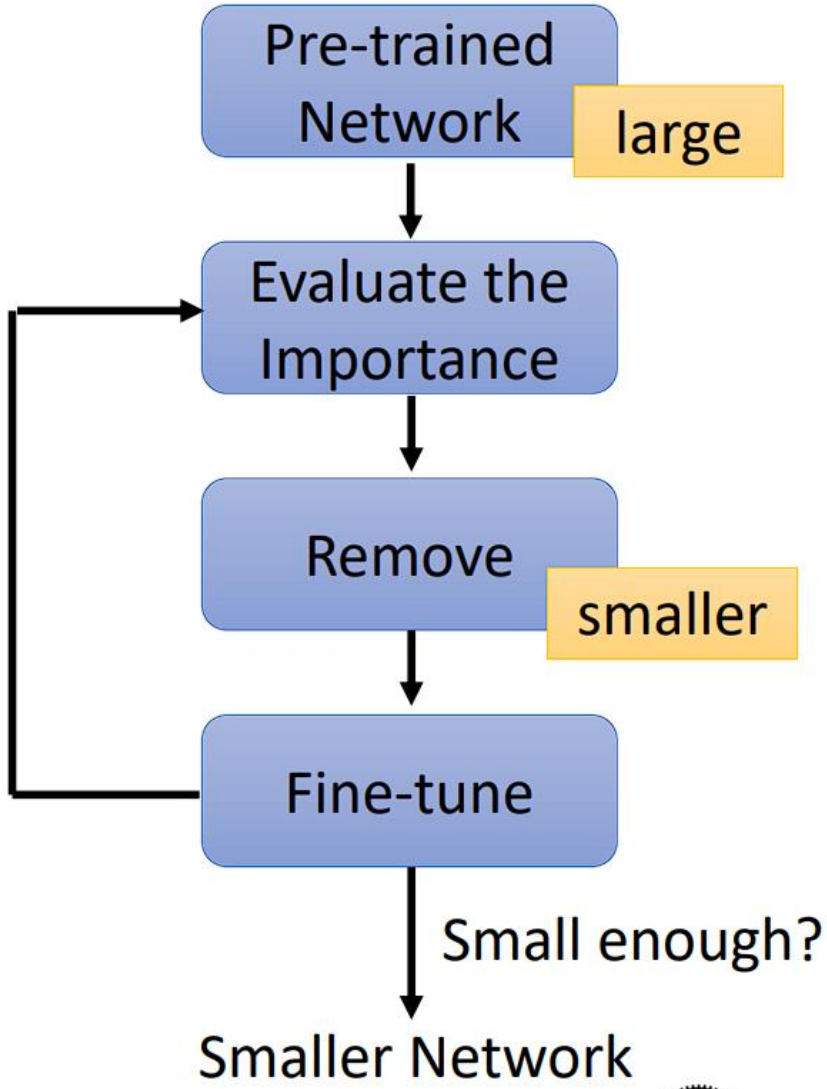
- Importance of a weight:

  absolute values, fisher-rao …

- Importance of a neuron:

  incident edge weights, the number of times it wasn't zero on a given data set ……

- After pruning, the accuracy will drop (hopefully minor)
- Fine-tuning on training data again
- Don't prune too much at once, or the network won't recover.




The slides are modified from the slides made by Prof. Hung-Yi Lee in NTU.

# Problem Description

BASIC LAB

- Dataset: Billnum
- Input: Fine-tuned T5-small (full parameter)
- Output: compressed model
- Constraint:
  - Ratio of non-zero parameter $\leq 0.3$
  - ROGUE-Lsum $\geq 0.1600$ (updated before **04/27 00:00**)
  - Do not use any test data or external data to finetune the model.

# Grading Policy

BASIC LAB

## Model Compression (total: 100%)

- Requirement :
  - Kaggle: team name must follow rule ([Student_id]_[Name] )
  - Report & Demo :
    - Report: Briefly introduce your work. (how to prune, use which package…)
    - TA could execute your code with provided README file.
- Kaggle Competition (100%)
  - Constrain: ratio of non-zero parameter ≤ 0.3
  - 80%: accuracy ≥ baseline benchmark
  - 20%: leaderboard ranking

20% scores are linearly distributed based on participants' rankings. The top-ranked participant receives the maximum points, and the lowest-ranked gets the minimum, with points evenly spread across ranks.

# Kaggle Competition

BASIC LAB

- Invitation Link:
  - https://www.kaggle.com/t/bb3b715e95d844c289426e638fa77445

- Timeline:
  - 4/27 00:00 Competition Starts
  - 5/14 23:59 Competition Finished & E3 Submission Deadline
- Team Name:
  - [Student_id]_[Name] (Requirement: [Student_id] must be correct)
  - Example: 0240309_佐藤和真(Using a nickname is also acceptable)

# Kaggle Competition

BASIC LAB

- Example submission
  - Two Column: “ID”,”Predict”
  - Make sure you turn off random setting for test data loader.

full_prediction.csv

T5-summarization > full_prediction.csv

```
"ID","Predict"
"0","Amends the Water Resources Development Act of 1992 to authorize the provision of an alternative water supply"
"1","Federal Forage Fee Act of 1993 - Requires all grazing operations conducted"
"2","Merchant Marine of World War II Congressional Gold Medal Act - Authorizes the Speaker of"
"3","Small Business Tax Modernization Act of 2004 - Amends the Internal Revenue Code to:"
"4","Fair Access to Investment Research Act of 2016 This bill directs the Securities and Exchange Commission (S"
"5","Prescription Drug Monitoring Act of 2016 This bill requires each covered state to require: (1) each prescri"
"6","Strategic Gasoline and Fuel Reserve Act of 2005 - Amends the Energy Policy and"
"7","Special Agent Scott K. Carey Public Safety Officer Benefits Enhancement Act - Amend"
"8","Promoting Financial Literacy and Economic Opportunity Act of 2015 This bill amends the Internal Revenue Code"
"9","Amends the Tariff Act of 1930 to prohibit a state. through its Attorney General"
"10","Outer Continental Shelf Revenue Sharing Act of 2005 - Amends the Outer Continental Shelf"
```



NYCU

# Special Rules

BASIC LAB

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 for compression is prohibited. TA will check the dataloader and execute your code.
5. Using pre-trained models created by others as the final result is prohibited. Please train your own model from .
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.

# Special Rules

BASIC LAB

1. Team Name Format:
   - Everyone should rename your team using the format: [student_id]_[name].
2. Submission Restrictions: Do not submit any results from the full model.
   - First violation: Your HW4 final score will be reduced by 50%. Please contact the TA as soon as possible to remove your incorrect submission.
   - Second violation: Your HW4 final score will be set to 0.
3. Model Validation:
   The TA will check that the ratio of non-zero parameters is ≤ 0.3 after the competition.
   If you accidentally submit results from an invalid model, you have only one opportunity to request the TA to remove it from the Kaggle leaderboard.
   - First time: There will be no penalty. Please contact the TA as soon as possible to remove your incorrect submission.
   - Second time: Your HW4 final score will be set to 0.
4. Additional Instructions:

- The TA will provide additional clarifications as needed.

# Demo Platform

BASIC LAB

- OS: Ubuntu Server 22.04
- CPU: Intel i7-8700
- GPU: RTX 4070 (12GB) *1
- Python 3.9
- CUDA: 12.1
- Framework: PyTorch 2.2.1

# Reminder

- Since the summary contains commas, which can conflict with the CSV format, please utilize the escape_special_characters function from the tutorial Jupyter notebook to address this issue.
- Since the Kaggle Competition Platform does not directly support the ROUGE-Lsum metric, the evaluation system will use a custom metric and code provided by the TA, which relies solely on built-in Python functions.
  Students are encouraged to perform self-evaluations using the calculate_lcs and score functions available in the tutorial before submitting their predictions.

# E3 Submission

BASIC LAB

- Two File:
  1. Report:
     <pdf file> HW4_[student_ID].pdf
     - Example: “HW4_9900940_曾偉倫.pdf”
  2. Code:
     <zip file> HW_[student_ID] _曾偉倫.zip
     - Example: “HW4_9900940_曾偉倫.zip”
     - Please make sure your submission contains the following items:
       1) All the code you used for training and testing (.py / .ipynb)
       2) The whole final weights folder used for testing
       3) A README file explaining how to execute your code (e.g., in txt or md format)
       4) Example: “HW4_9900940_曾偉倫.zip”
          --- prining_T5.ipynb / prining_T5.py
          --- pruned_model/
          --- README.txt /.pdf /.docx ...

NYCU

# Resource

BASIC LAB

- Colab: https://colab.research.google.com/
- T5: https://huggingface.co/google-t5/t5-small
- Summarization: https://huggingface.co/docs/transformers/tasks/summarization