

Data Science HW #4

Model Compression for LLM

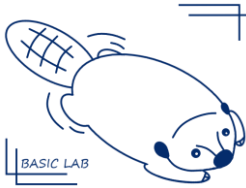
TA: 曾偉倫

Email: wlt seng.ee06@nycu.edu.tw

Table of Contents

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

Introduction



BASIC LAB

Model Compression Goals

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



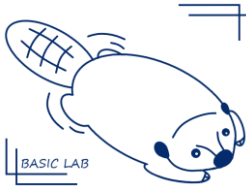
Tradeoff for Network Compressions

Model Performance



Compression Rate

Sentiment Analysis



BASIC LAB

- ChatGPT Demo

You

Please do sentiment analysis, the following text should belong to positive / neutral / negative

However , the suspect stole his burgundy Nissan Altima .

ChatGPT

The sentiment analysis for the given text "However, the suspect stole his burgundy Nissan Altima." is as follows:

- Sentiment: Negative

This analysis indicates a negative sentiment due to the occurrence of the word "stole," which implies a criminal act. If you need further clarification or have additional text for analysis, feel free to provide it!



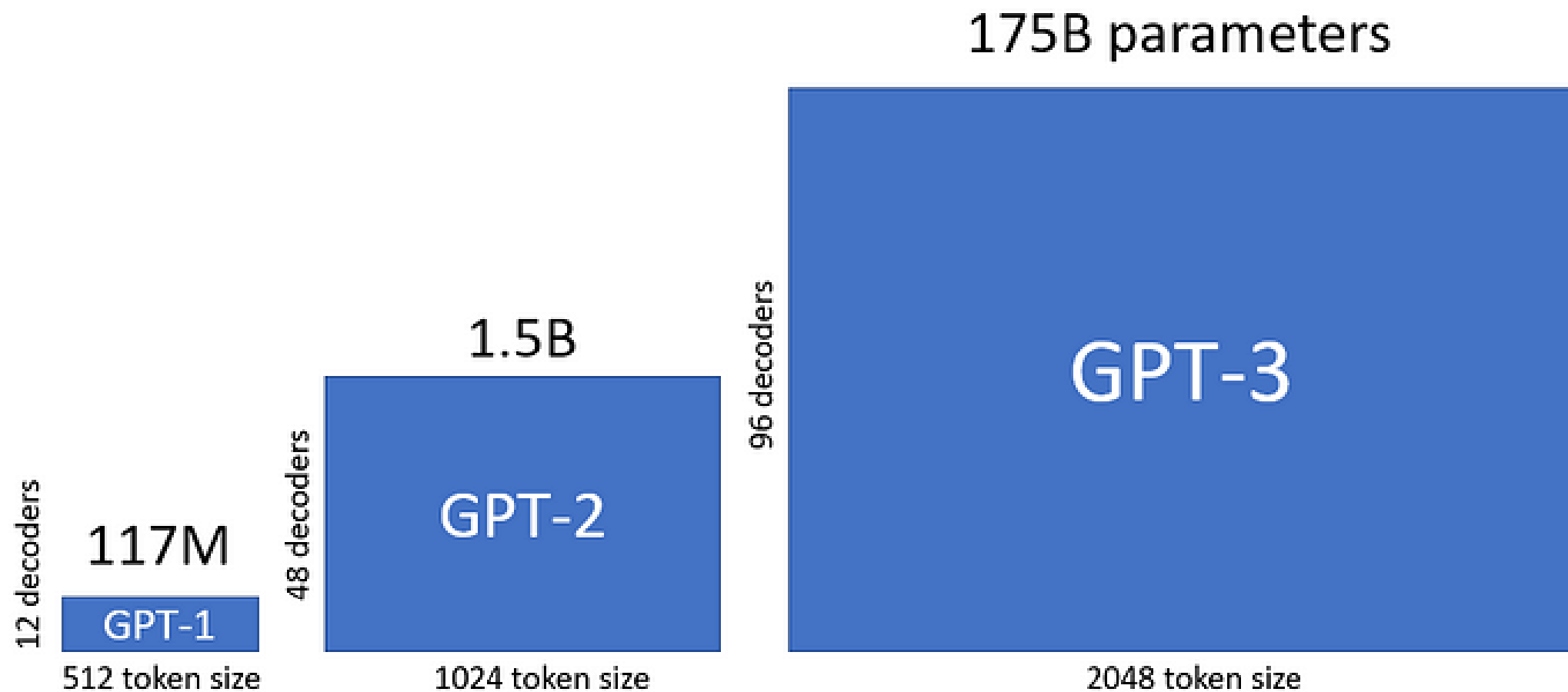


Fig. Evolution of GPT-1, GPT-2 and GPT-3.

(ref: <https://medium.com/@YanAlx/step-by-step-into-gpt-70bc4a5d8714>)

Problem Description

- Dataset: Sentiment Analysis
- Input: GPT-2 pre-trained model & training data
- Output: Fine-tuned GPT-2 model
- Constrain:
 - Model size: 245 MB (details in the later slide)
 - Accuracy ≥ 0.592
 - **DO NOT USE ANY TEST DATA, EXTERNAL DATA, EXTERNAL MODEL**

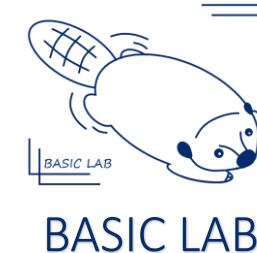
Problem Description



- Dataset
 - Sentiment Analysis
 - File:
 - public.csv
 - private_release.csv



public.csv



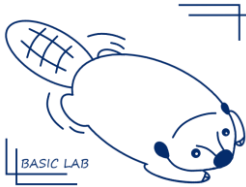
- sentiment,text

	sentiment	text
1	positive	it part of the company 's strategy in the region of central and eastern Europe , " said Jozef Misurec , the director of Ruukki Slovakia .
2	positive	Ruukki has signed a contract to deliver and install the steel structures for a bridge over the Kyrnsalmi strait in Savonlinna , Finland .
3	positive	The production capacity can be tripled without any major additional investments , the company said .
4	positive	Commission income rose by 25.7 % to EUR 16.1 mn from EUR 12.8 mn in 2004 .
5	positive	These measures are expected to produce annual cost savings of EUR 1.5 million starting in 2007 .
556	neutral	Rimi supermarket is the key customer in Magistral center .
557	neutral	" We know there are some of our own people out there . "
558	neutral	Results are expected late in 2006 .
559	neutral	m Web site will display submissions for the expanded Broadband Emmy Awards as part of an effort to identify aspiring video artists .
560	neutral	en new Finnish nuclear power stations go into operation and the large hydro powered stations of northern Europe have a good year .
1016	negative	Pretax profit totaled EUR 4.9 mn compared to EUR 5.2 mn in the first quarter of 2005 .
1017	negative	However , the suspect stole his burgundy Nissan Altima .
1018	negative	Operating profit margin was 8.3 % , compared to 11.8 % a year earlier .
1019	negative	Ramirent made 18 million kroons EUR 1.15 mln loss last year ; the year before the company was 7.3 million kroons in the black .
1020	negative	& Paper Machinery unit was over EUR 3mn in September 2007 - August 2008 , compared to a profit of EUR 3.7 mn a year earlier .

kaggle.csv

- text

	text
1	nd reports its operating profit fell to EUR 6.1 mn in the second quarter of 2008 from EUR 7.5 mn in the second quarter of 2007 .
2	total of 8.7 mn mobile devices were sold in the area , representing an increase of 32 % from the corresponding period in 2009 .
3	osphodiesterase 10 inhibitor program since its start in late 2006 and has the sole authority to develop discovered compounds .
4	icture , communications and networks service platforms , as well as professional services , to operators and service providers .
5	orkers ' Union SEL plans to hasten its collective bargaining with a two-day strike that would begin on 7 April 2010 , in Finland .
6	erve a broad range of customers , including construction and process industries , shipyards , the public sector and households .
7	truction of three ferryboats with an option for two more units , further confirmed at a total value of 500 mln euro (\$ 645.8 mln) .
8	There has been some recovery of the base metals prices and increased demand for metals in China , however .
9	Raute Corporation (OMX Helsinki : RUTAV) issued on Tuesday (23 September) a profit warning for the financial year 2008 .
10	Mississippi and the region known as the logistics center of the United States , " said Gray Swoope , executive director of MDA .
11	orted on Monday (18 August) an operating loss of EUR3 .3 m on net sales of EUR33 .9 m for the period January-June 2008 .
12	ata Solution Oyj reported a net loss of 11.7 mln euro \$ 17.2 mln for 2007 versus a net profit of 2.5 mln euro \$ 3.7 mln for 2006 .
13	otiations concern personnel of Cencorp Corporation and Singulase Oy as whole in Finland and in Sweden , the company said .
14	of Boomeranger Boats Oy is in line with the Done Solutions strategy of developing its business through company acquisitions .
15	olidated net operating profit increased from EUR 4.8 mn in the first quarter of 2005 to EUR 6.4 mn in the first quarter of 2006 .



BASIC LAB

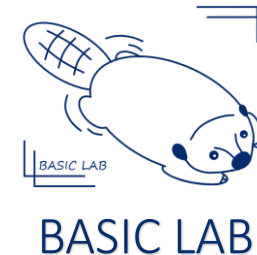
Problem Description

- Pre-trained Model
 - Please fine-tune GPT-2 from “transformers” package

```
from transformers import GPT2Tokenizer, GPT2LMHeadModel

model = GPT2LMHeadModel.from_pretrained('gpt2', device_map=device, ...)
tokenizer = GPT2Tokenizer.from_pretrained('gpt2')
```

Baseline

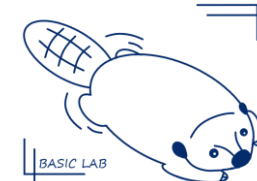


- GPT-2 configuration

- Bnb_config: `compute_dtype = getattr(torch, "float16")`
`bnb_config = BitsAndBytesConfig(load_in_4bit=True,
bnb_4bit_quant_type="nf4",
bnb_4bit_compute_dtype=compute_dtype,
bnb_4bit_use_double_quant=True,)`

- LoraConfig: `lora_alpha=16, lora_dropout=0.1, r=64,
bias="none", target_modules="all-linear",
task_type="CAUSAL_LM",`

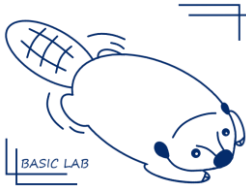
Baseline



BASIC LAB

```
training_arguments = TrainingArguments(  
    output_dir=output_dir,  
    num_train_epochs=3,  
    per_device_train_batch_size=1,  
    gradient_accumulation_steps=8,  
    gradient_checkpointing=True,  
    optim="paged_adamw_32bit",  
    save_steps=0,  
    logging_steps=25,  
    learning_rate=2e-4,  
    weight_decay=0.001,  
    fp16=True,  
    bf16=False,  
    max_grad_norm=0.3,  
    max_steps=-1,  
    warmup_ratio=0.03,  
    group_by_length=True,  
    lr_scheduler_type="cosine",  
    report_to="tensorboard",  
    evaluation_strategy="epoch"  
)  
  
# directory to save and repository id  
# number of training epochs  
# batch size per device during training  
# number of steps before performing a backward/update pass  
# use gradient checkpointing to save memory  
  
# Log every 10 steps  
# Learning rate, based on QLoRA paper  
  
# max gradient norm based on QLoRA paper  
  
# warmup ratio based on QLoRA paper  
  
# use cosine learning rate scheduler  
# report metrics to tensorboard  
# save checkpoint every epoch
```

```
trainer = SFTTrainer(  
    model=model,  
    args=training_arguments,  
    train_dataset=train_data,  
    peft_config=peft_config,  
    eval_dataset=eval_data,  
    dataset_text_field="text",  
    tokenizer=tokenizer,  
    max_seq_length=1024,  
    packing=False,  
    dataset_kwargs={  
        "add_special_tokens": False,  
        "append_concat_token": False,  
    }  
)
```



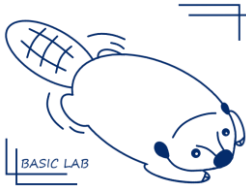
BASIC LAB

How to check model size?

- TA will provide this example code in HW supplementary.
 - Check the size of entire model folder ≤ 245 MB

```
1 finetuned_model = "YOUR MODEL FOLDER PATH"
2
3 def get_folder_size(folder_path):
4     total_size = 0
5     for dirpath, dirnames, filenames in os.walk(folder_path):
6         for filename in filenames:
7             file_path = os.path.join(dirpath, filename)
8             total_size += os.path.getsize(file_path)
9     return total_size / (1024*1024) # Convert bytes to megabytes
10
11
12 folder_size_MB = get_folder_size(finetuned_model)
13 print("Folder size:", folder_size_MB, "MB")
14
15
```

Grading Policy



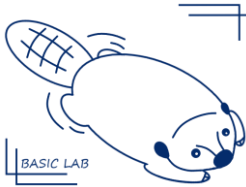
BASIC LAB

Model Compression for LLM (total: 100%)

- Kaggle Competition (75%)
- Report (20%)
- Submission Correctness & Demo (5%)



Grading Policy



BASIC LAB

Model Compression for LLM(total: 100%)

- Kaggle Competition (45%+30%)
 - Constrain: Model full size (folder size) ≤ 245 MB
 - 45%: Accuracy \geq baseline benchmark
 - 30%: Final leaderboard ranking

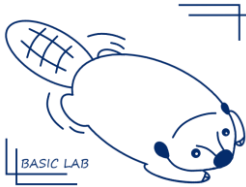


Kaggle Competition



- Invitation Link:
 - [<Under construction>](#)
- Submission details will be released on competition site.
- A maximum of 5 submissions per day is allowed on Kaggle.
- Timeline:
 - 4/17 00:00 Kaggle Competition Start
 - 5/3 23:59 Kaggle Competition Finished
 - 5/7 23:59 E3 Submission Deadline

Kaggle Competition

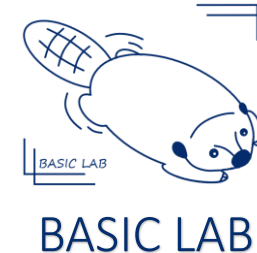


BASIC LAB

- Example submission
 - Only one column, no need id
 - Make sure you turn off random setting for test data loader.
 - No need column name (first row is the first prediction)
 - Sentiment mapping
 - 'positive': 2
 - 'neutral': 1
 - 'negative': 0

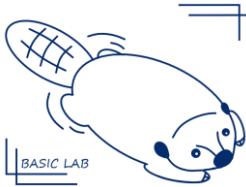
1
2
0
2
0
0
1
0
2
0
1
1
0
2
0
1

Grading Policy



- Report (20%)
 - Model size (5%)
 - Brief Explanation of Compression Methods (15%)
 - File name: HW4_[student_ID]_[Name].pdf
(Example: HW4_71000007_葉宜麗.pdf)
 - Methods you used (compression config, training arguments, ...)
 - Reference (any reference from GitHub, Hugging Face, ...)
 - ≤ 200 words, remember to write your name & student id in the report.
- Demo (5%)
 - TA will execute your code and reproduce the results.
 - DO NOT use

Special Rules

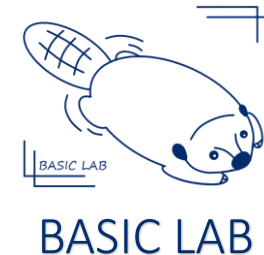


BASIC LAB

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 fine-tune GPT-2 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.

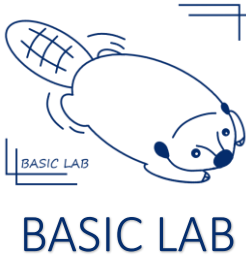
Violation of any of the above rules will result in a score of 0 for this assignment.

Demo Platform



- OS: Ubuntu Server 20.04
- CPU: Intel i7-8700 (will set num_worker=8)
- GPU: RTX 4070 (12GB) *1
- Python 3.9
- CUDA: 12.1
- Framework: PyTorch 2.2.1

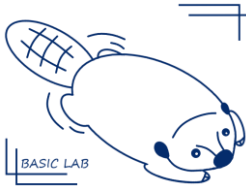
E3 Submission



- Two File:
 1. <pdf file> HW4_[student_ID].pdf
 - Example: “HW4_71000007_葉宜麗.pdf”
 2. <zip file> HW_[student_ID]_葉宜麗.zip
 - Example: “hw2_311000123_葉宜麗.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: “hw2_311000123_葉宜麗.zip”
 - compression.ipynb
 - merged_model/
 - README.txt

Resource

- Colab: <https://colab.research.google.com/>
- GPT2: <https://huggingface.co/openai-community/gpt2>
- Kaggle Tutorial:
[Fine-Tuning with Llama 2 + QLoRA \(kaggle.com\)](https://www.kaggle.com/learn/llm2)



BASIC LAB

推薦工具

- Colab: <https://colab.research.google.com/>
- 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 \(hostinger.com\)](#)