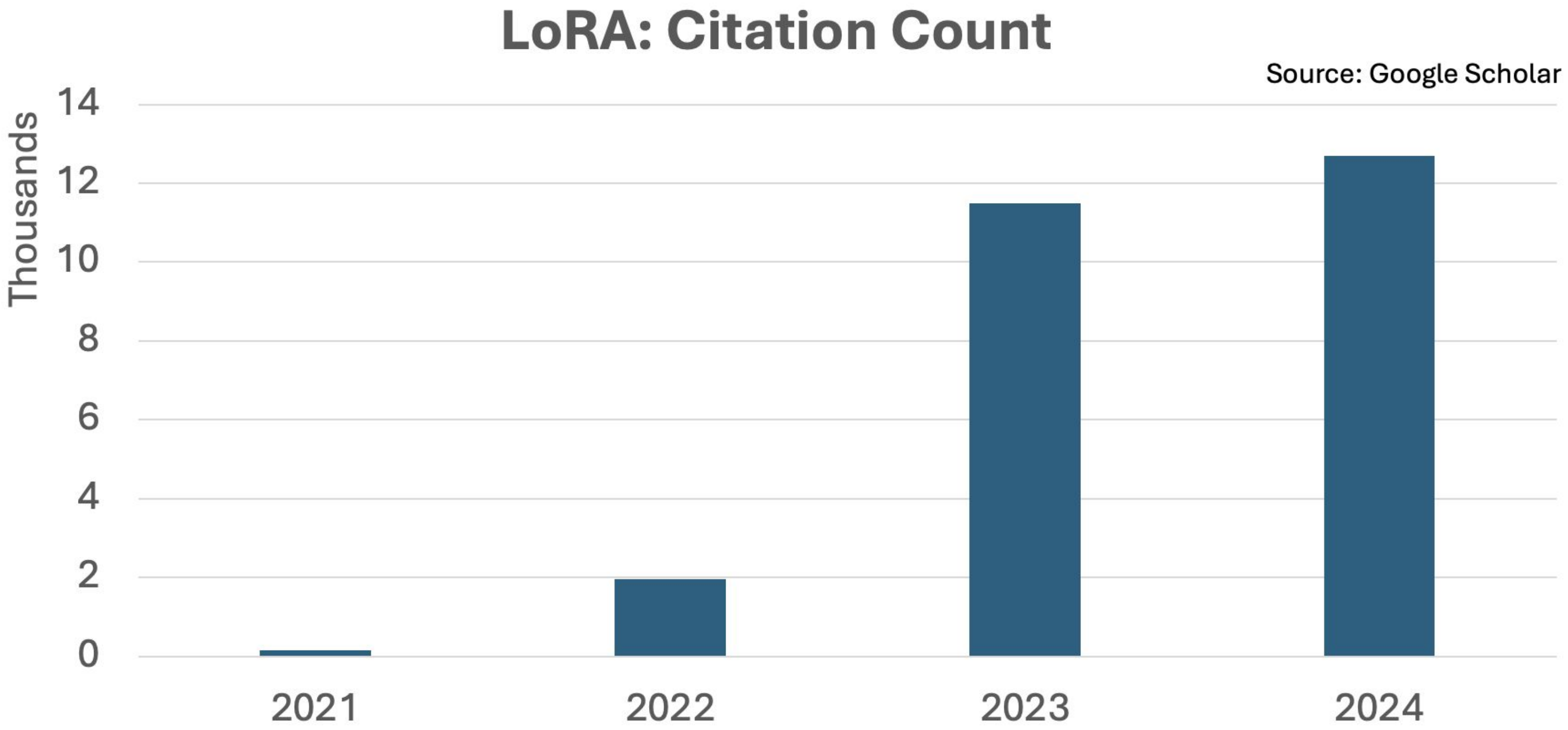
Low-Rank Adaption of Large Language Model (LoRA)

by Edward J. Hu et al.

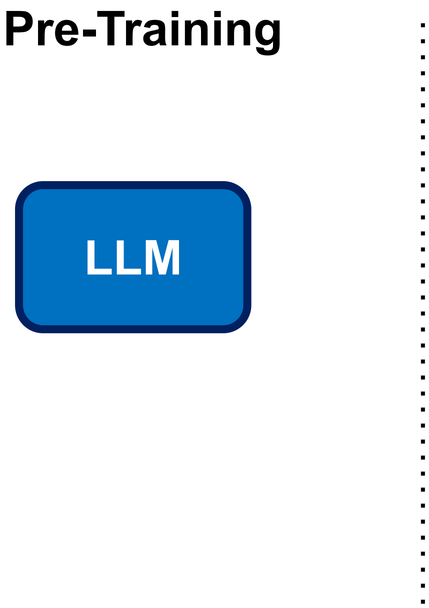
Group 5

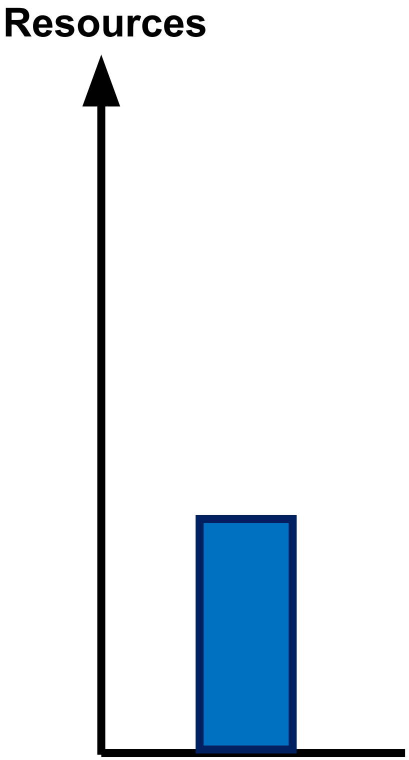
Yang, Gugler, Gube

Is it relevant?

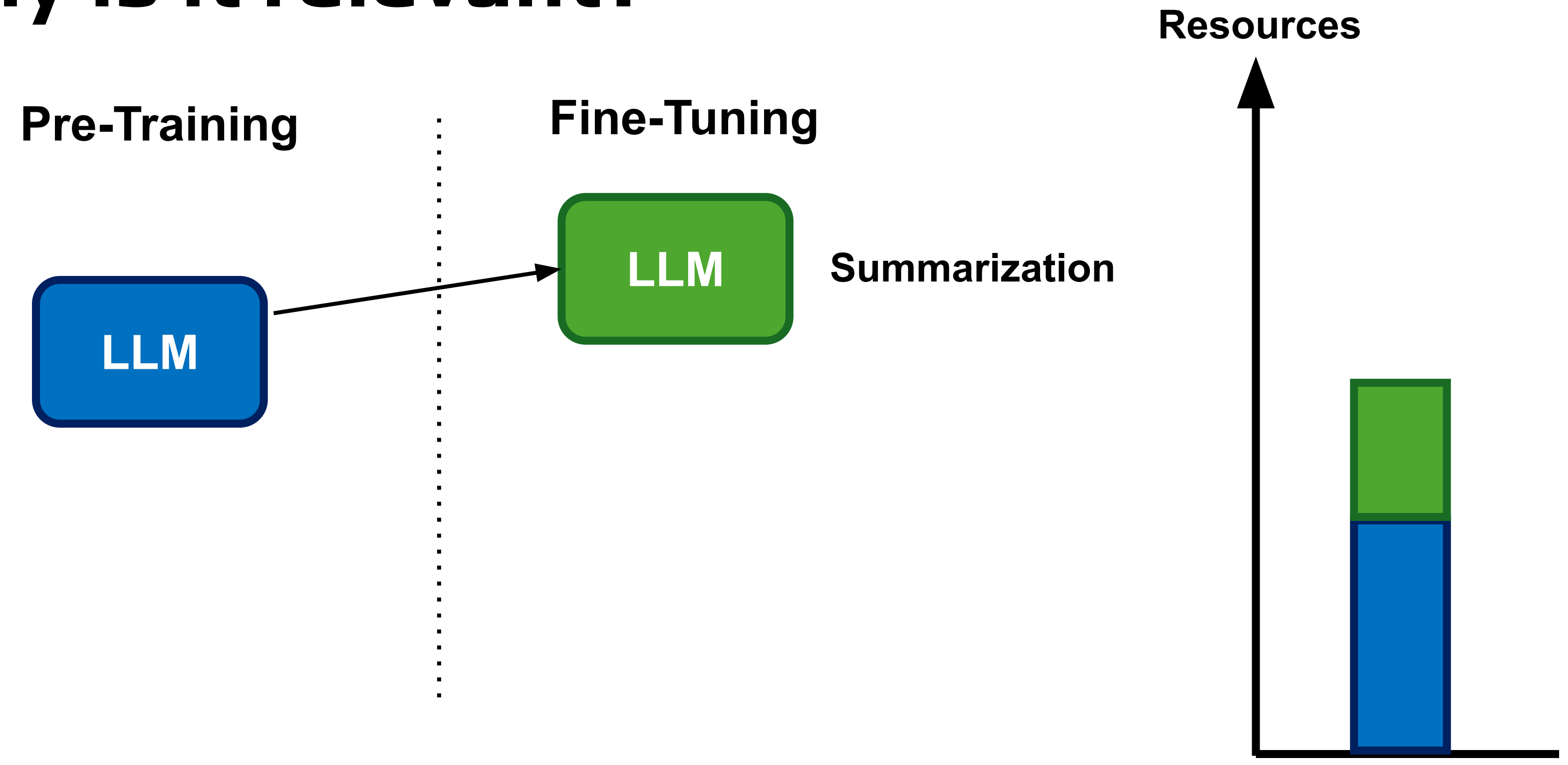


Why is it relevant?

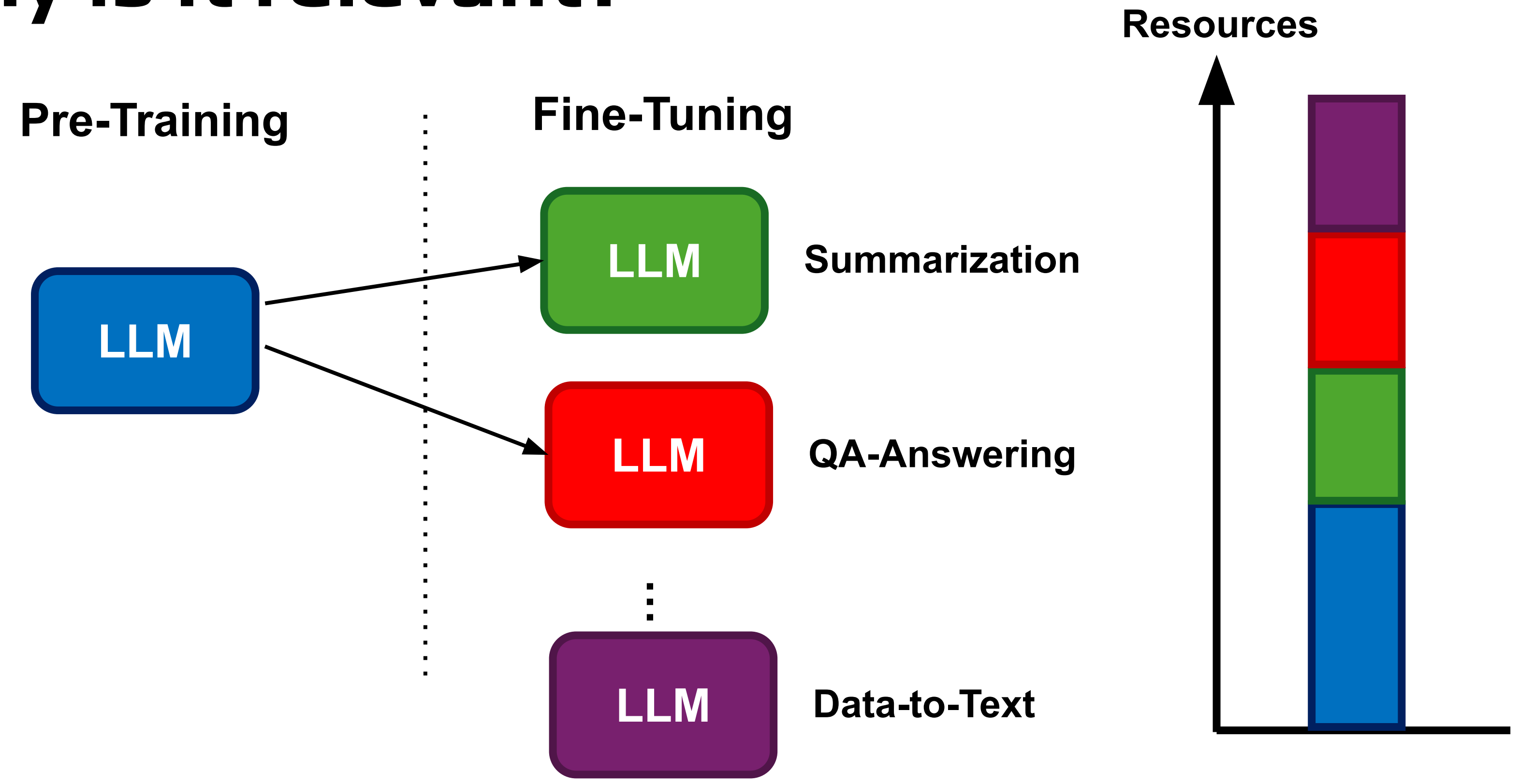




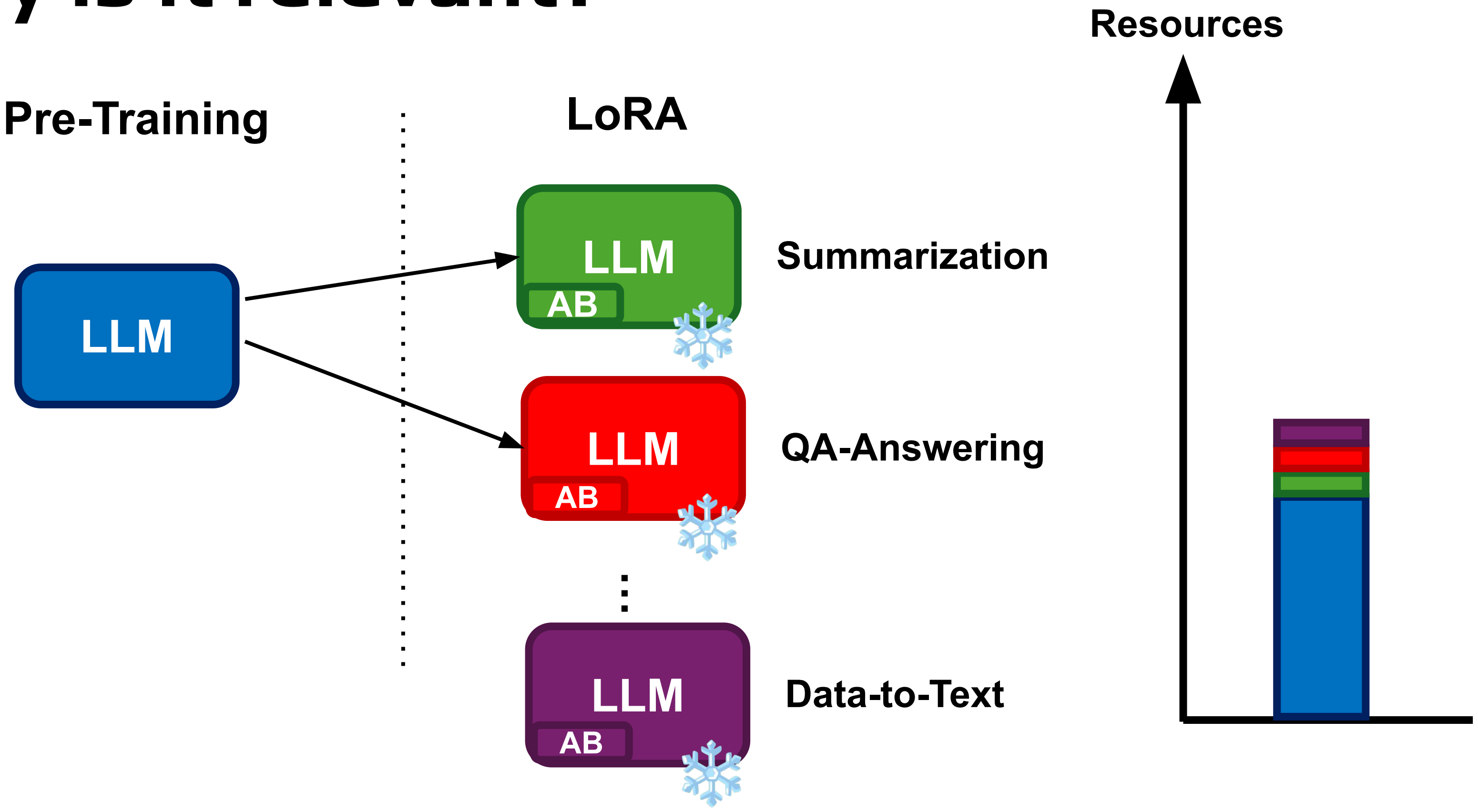
Why is it relevant?



Why is it relevant?



Why is it relevant?

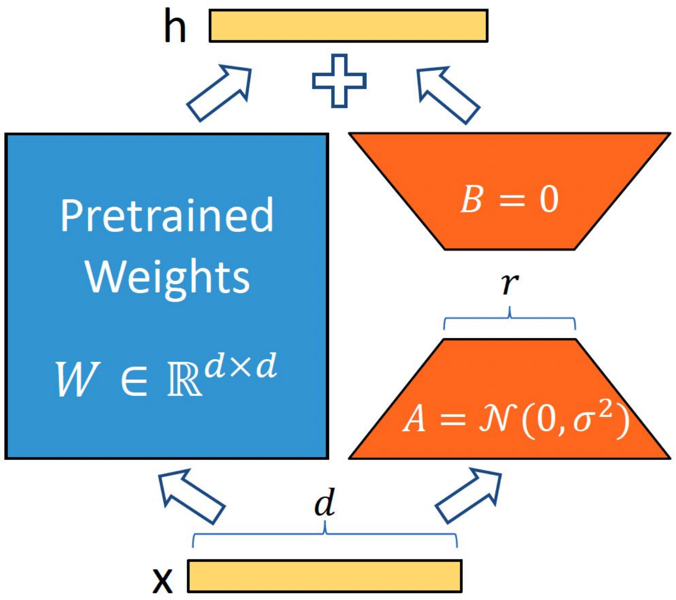


Core Mechanisms

- Adaption to specific task: low intrinsic dimension [Aghajanyan et al., 2020]
- Weight adaption ΔW is also low
- Idea:
    - Freeze Model params W
    - Replace ΔW with AB, two low rank matrices

$$\boxed{\boldsymbol{\Delta W} \approx \boldsymbol{BA}}$$

$$W \in \mathbb{R}^{d\times k}, A \in \mathbb{R}^{r\times k}, B \in \mathbb{R}^{d\times r}, r \ll d$$

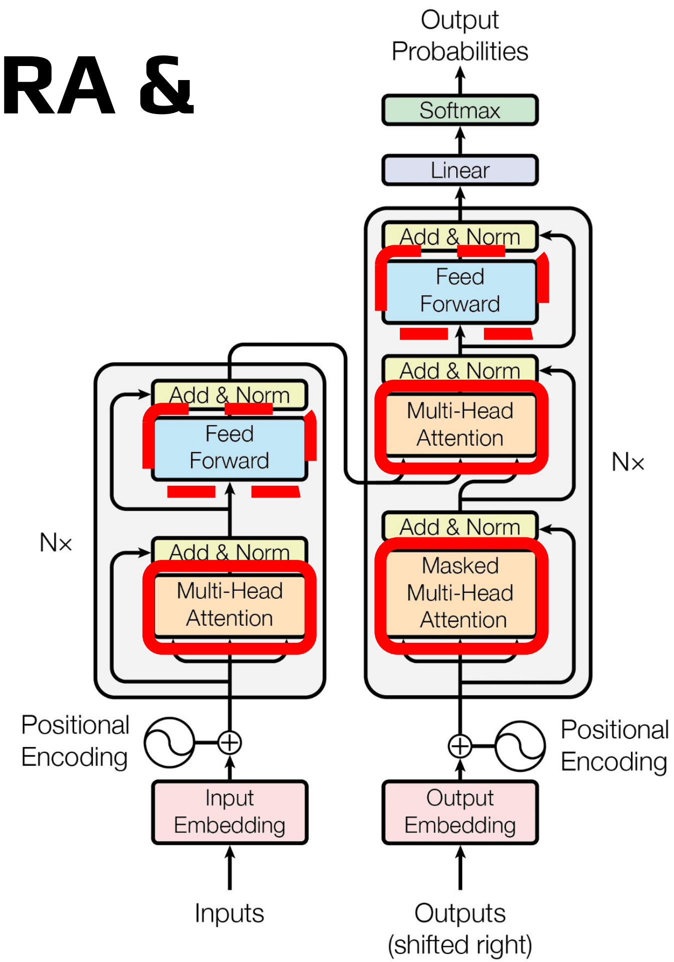


Core Mechanisms: LoRA & Transformers

- LoRA Operation Points:
  - Attention modules: W_q, W_k, W_v, W_o

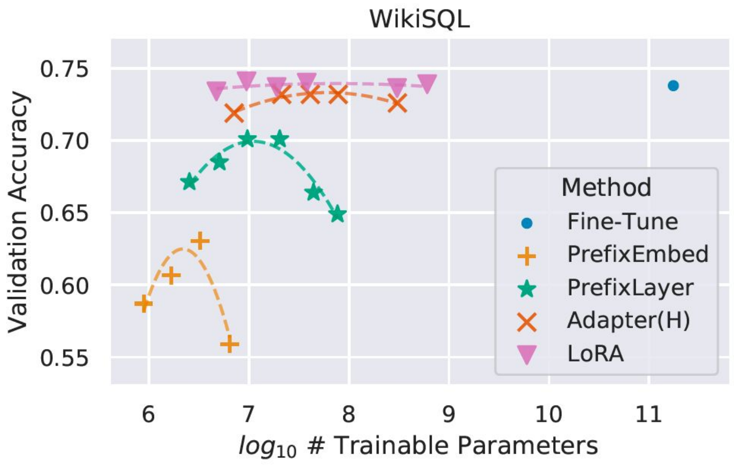
$$softmax(\frac{QK^T}{\sqrt{d_k}})VW_o$$

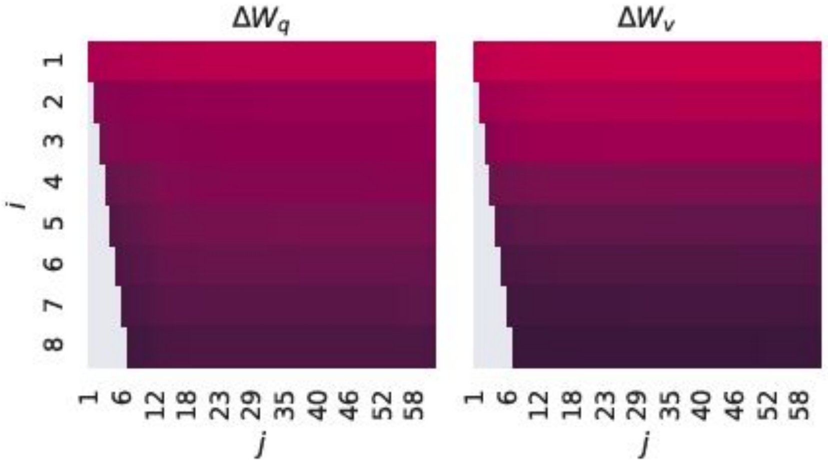
  - (Feed Forward Layers)



Evaluation

- **Models:** RoBERTa, DeBERTa, GPT-2, GPT-3
- **Competitive Techniques**: Fine-Tuning, Adapter, Prefix-Tuning
- **Experiments:** Performance, “Good Rank“, Subspace similarity





RoBERTa

RoBERTa Experimental Setup

- Base Model: RoBERTa-base (125M parameters)
- Libraries: Hugging Face Transformers, PEFT (LoRA integration)
- Dataset: 8 dataset from **GLUE** benchmark

Domains:

- News / Movie reviews / Wikipedia / Books / Misc.

Classification Tasks:

- Acceptability
- Sentiment
- Paraphrase
- Sentence similarity
- Question Answering
- Natural Language Inference

RoBERTa Implementation Details

- Used same parameters from paper:

Method	Dataset	MNLI	SST-2	MRPC	CoLA	QNLI	QQP	RTE	STS-B
	Optimizer	AdamW							
	Warmup Ratio	0.06							
	LR Schedule	Linear							
RoBERTa base LoRA	Batch Size	16	16	16	32	32	16	32	16
	# Epochs	30	60	30	80	25	25	80	40
	Learning Rate	5E-04	5E-04	4E-04	4E-04	4E-04	5E-04	5E-04	4E-04
	LoRA Config.	$r_q = r_v = 8$							
	LoRA α	8							
	Max Seq. Len.	512							

RoBERTa Implementation Details

- Used same parameters from paper:

Method	Dataset	MNLI	SST-2	MRPC	CoLA	QNLI	QQP	RTE	STS-B
	Optimizer	AdamW							
	Warmup Ratio	0.06							
	LR Schedule	Linear							
RoBERTa base LoRA	Batch Size	16	16	16	32	32	16	32	16
	# Epochs	~~30~~ 8	60	30	80	25	~~25~~ 12	80	40
	Learning Rate	5E-04	5E-04	4E-04	4E-04	4E-04	5E-04	5E-04	4E-04
	LoRA Config.	$r_q = r_v = 8$							
	LoRA α	8							
	Max Seq. Len.	512							

- Reduced #epochs for MNLI, and QQP datasets due to time constraints

RoBERTa Results

- Results:

Dataset	Reference Implementation	Our Implementation	Hyperparameter adjustments
MNLI	87.5	84.47	✓# Epochs: reduced from 30 to 8
SST-2	95.1	94.69	-
MRPC	89.7	**73.45**[1] / 86.55[2]	-
CoLA	63.4	60.25	-
QNLI	93.3	**62.47**	-
QQP	90.8	89.68	✓# Epochs: reduced from 25 to 12
RTE	86.6	**54.87**[1] / 74.73[2]	-
STS-B	91.5	**70.82**[1] / 90.82[2]	-
Avg.	87.2	73.84[1] / 80.46[2]	

1 = LoRA adapter initialized to best MNLI checkpoint, 2 = LoRA initialized to Gaussian Random Field

RoBERTa Results

- Results:

  Able to reproduce results for **4 datasets**
  - MNLI, SST-2, CoLa, QQP

  Negative impact by using LoRA adapters trained on MNLI:
  - MRPC, RTE, STS-B

  Unable to reproduce:
  -QNLI

GPT-2

GPT-2 Experimental Setup

- Base Model: GPT-2-medium (345M parameters)
- Libraries: Hugging Face Transformers, PEFT (LoRA integration)
- Dataset: E2E NLG Challenge (restaurant domain)
  - 42,000 training examples, 4,600 test/validation examples
  - Meaning representations (MR) paired with natural language references

```
name[The Eagle], food[English], price[moderate], customer rating[3/5]
```

⬇

```
The Eagle is a moderately priced English restaurant with a 3/5 customer rating.
```

GPT-2 Code Snippet: LoRA Configuration

```
lora_config = LoraConfig(
    r=4,
    lora_alpha=32,
    target_modules=["c_attn"],
    lora_dropout=0.1,
    init_lora_weights="gaussian",
    bias="none"
)
model = get_peft_model(model, lora_config)
```

GPT-2 Evaluation Metrics

- Metrics: BLEU, METEOR, ROUGE-L, NIST, CIDEr
- Generation Strategy: Beam search (width 10, length penalty 0.9)
- Results:
  - Comparable to original LoRA paper
  - Minor differences due to Python-based metrics

Metric	Reference Implementation	Our Implementation
BLEU	0.6603	0.6619
METEOR	0.8139	0.8183
ROUGE-L	0.6750	0.6530
NIST	7.2166	7.1026
CIDEr	2.1189	2.1522

GPT-2 Challenges and Solutions

- Padding Handling: Proper padding token configuration
  - Solved by configuring padding tokens and setting `padding_side='left'`
- Memory Management: Efficient batching and gradient accumulation
  - Utilizing transformer Data collators for dynamic batching
- Hyperparameters: Required checking reference implementation for missing details

Evaluation of Best Rank r

- Model: GPT-2 Medium
- Dataset: E2E

Rank r	val_loss	BLEU	NIST	METEOR	ROUGE_L	CIDEr
1	1.23	68.72	8.7215	0.4565	0.7052	2.4329
2	1.21	69.17	8.7413	0.4590	0.7052	2.4639
4	1.18	**70.38**	**8.8439**	**0.4689**	0.7186	**2.5349**
8	1.17	69.57	8.7457	0.4636	**0.7196**	2.5196
16	**1.16**	69.61	8.7483	0.4629	0.7177	2.4985
32	**1.16**	69.33	8.7736	0.4642	0.7105	2.5255
64	**1.16**	69.24	8.7174	0.4651	0.7180	2.5070
128	**1.16**	68.73	8.6718	0.4628	0.7127	2.5030
256	**1.16**	68.92	8.6982	0.4629	0.7128	2.5012
512	**1.16**	68.78	8.6857	0.4637	0.7128	2.5025
1024	1.17	69.37	8.7495	0.4659	0.7149	2.5090

Evaluation of Best Rank r

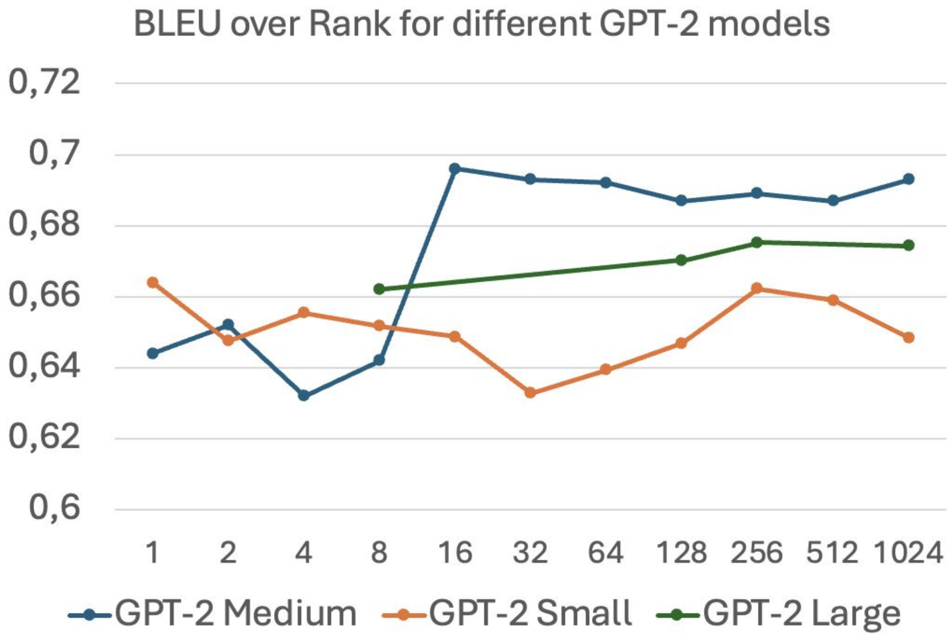
- Model: GPT-2 Medium
- Dataset: E2E

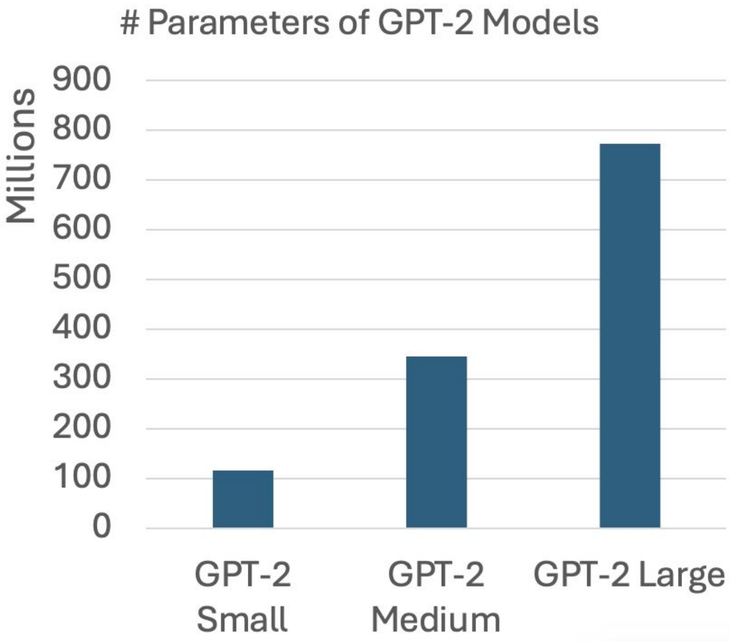
Observed Percentual Deviations

Rank r	BLEU (%)	NIST (%)	ROUGE (%)	CIDEr (%)
1	6.26	20.90	7.16	15.63
2	5.50	18.94	4.10	7.87
4	14.23	22.22	6.68	12.10
8	7.63	19.35	6.41	9.84
16	4.89	17.55	4.70	5.84
32	3.61	16.62	3.55	6.09
64	6.21	19.14	6.02	9.06
128	3.49	17.31	3.79	5.75
256	4.64	17.92	4.08	7.41
512	3.93	17.27	5.05	5.91
1024	3.32	16.92	4.11	7.44

Extension: Rank & Model Size

- Effects of Rank & Model Size on Performance





Extension: Rank & Task Difficulty

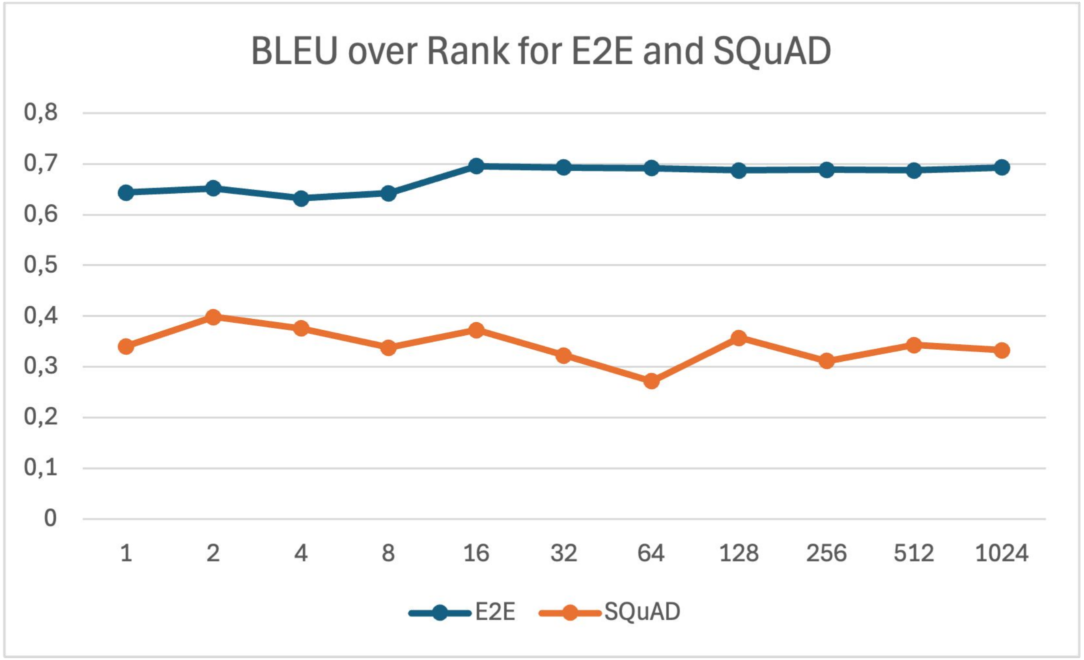
- Effects of Rank & Task Difficulty on Performance
- SQuAD: Extractive QA


```
"answers": {
    "answer_start": [
        1
    ],
    "text": [
        "This is a test text"
    ]
},
"context": "This is a test context.",
"id": "1",
"question": "Is this a test?",
"title": "train test"
```


<Q> Is this a test? <C> This is a test context. <A> This is a test text

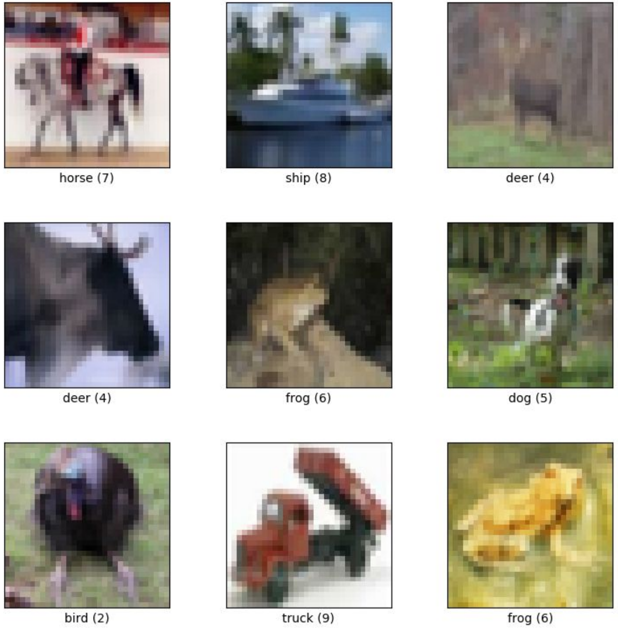
Extension: Rank & Task Difficulty

- SQuAD: easy task
- Correlation between difficulty and rank



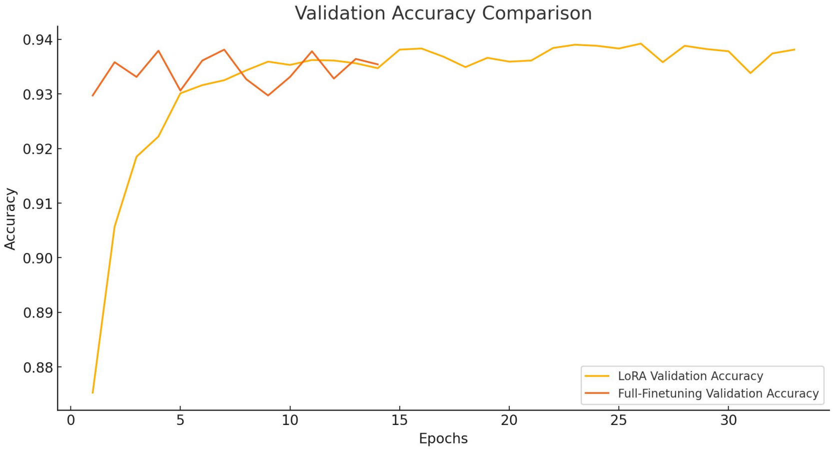
Ext. Image Classification Experimental Setup

- Dataset: CIFAR-10 (60,000 images, 10 classes)
- Model: ResNet-18 with LoRA adaptors
- Implementation:
  - LoRA rank (r): 14
  - LoRA alpha: 16
  - Dropout: 0.1
- Hyperparameters:
  - Learning rate: 1e-4
  - Batch size: 32
  - Early stopping: Patience of 7 epochs
- Optimizer: Adam with cross-entropy loss
- Data Augmentation: Resizing, normalization



Ext. Image Classification Results and Analysis

- Parameter Efficiency:
  - Trainable parameters reduced by 96.3% (11.7M → 435K)
- Performance:
  - Validation accuracy: 93.92%
  - Test accuracy: 93.50%
  - Comparable to standard fine-tuning (93.78%)



Ext. Image Classification Conclusion

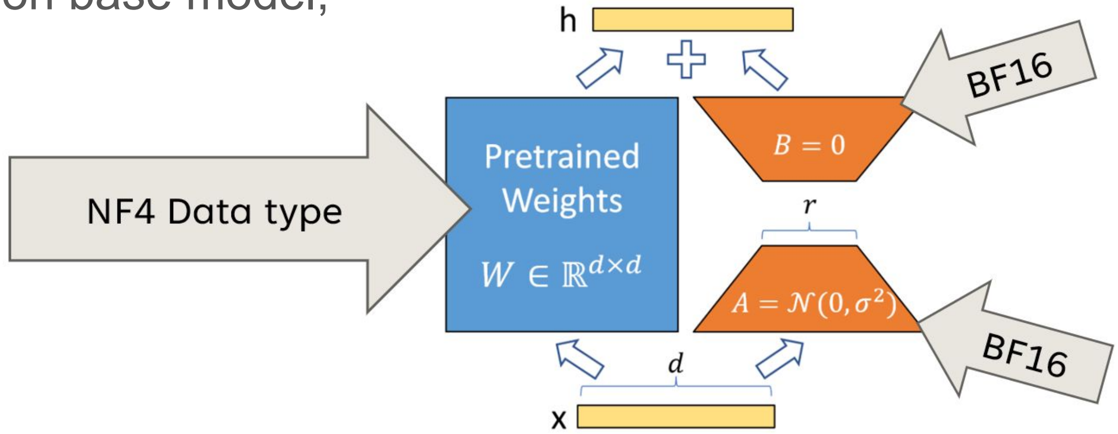
- Demonstrated LoRA's effectiveness in image classification.
- Achieved high accuracy with significantly fewer trainable parameters.
- Future Work:
  - Explore LoRA on larger vision models (e.g., ViT, Swin Transformers).
  - Investigate LoRA for object detection and segmentation tasks.

QLoRA

QLoRA Implementation Details

- Quantized LoRA
  Quantization only applied on base model, not LoRA adapters

- Contributions:
  - New NF4 datatype
  - Double Quantization
  - Paged Optimizers
- Reduces memory during training / inference



QLoRA

- New NF4 datatype:

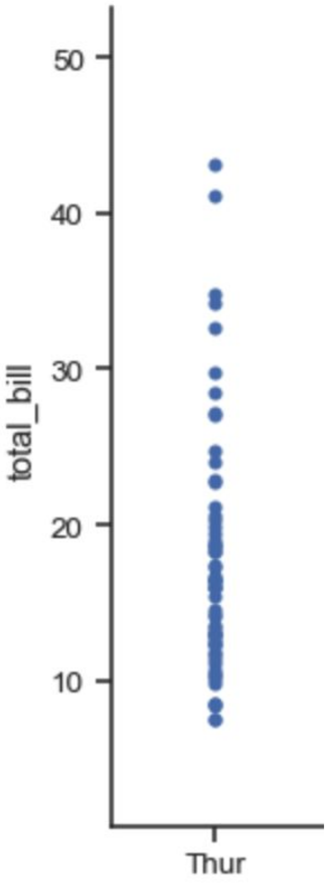
  NormalFloat - 4bit -> 2^4 -> 16 values

QLoRA

- New NF4 datatype:

  NormalFloat - 4bit -> 2^4 -> 16 values

- Let’s assume we have the following dataset:

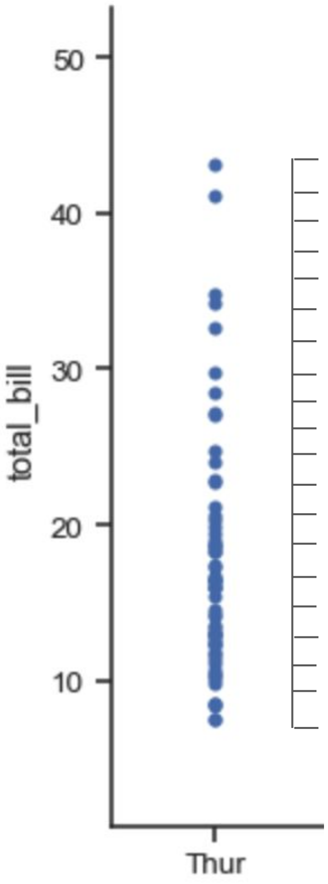


QLoRA

- New NF4 datatype:

  NormalFloat - 4bit -> 2^4 -> 16 values

- Let’s assume we have the following dataset:
  1) We then partition the scale into 16 equally sized parts from *max* and *min*

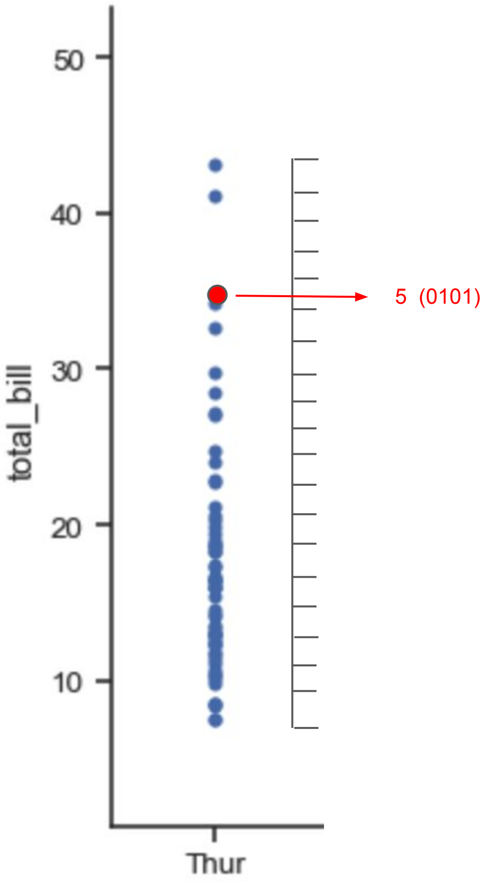


QLoRA

- New NF4 datatype:

  NormalFloat - 4bit -> 2^4 -> 16 values

- Let's assume we have the following dataset:
  1) We then partition the scale into 16 equally sized parts from *max* and *min*
  2) We map the value to the nearest NF value

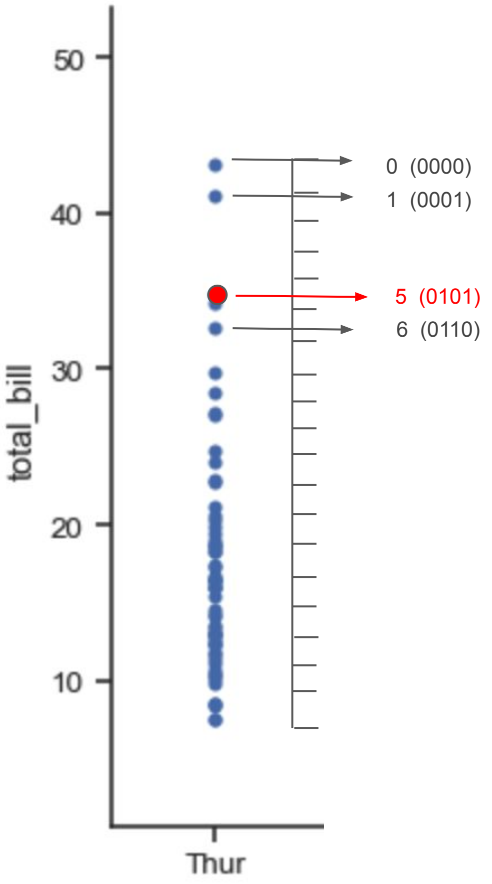


QLoRA

- New NF4 datatype:

  NormalFloat - 4bit -> 2^4 -> 16 values

- Let’s assume we have the following dataset:
  1) We then partition the scale into 16 equally sized parts from *max* and *min*
  2) We map the value to the nearest NF value

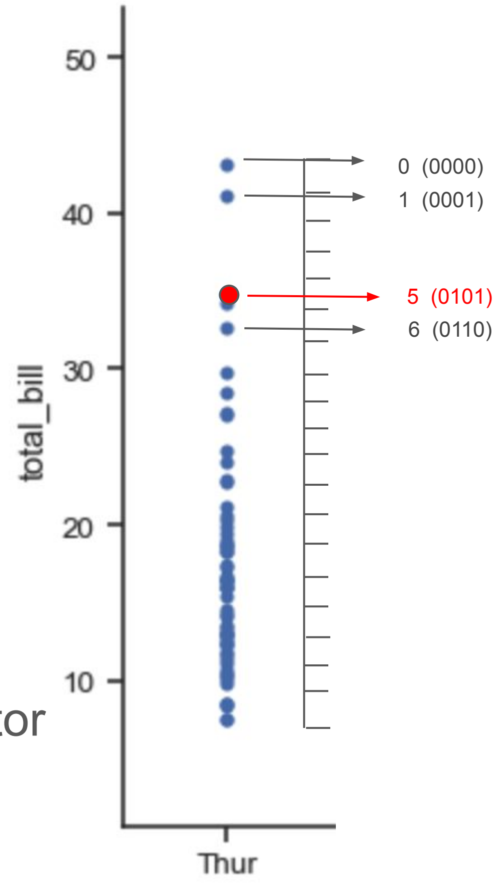


QLoRA

- New NF4 datatype:

  NormalFloat - 4bit -> 2^4 -> 16 values

- Let’s assume we have the following dataset:
  1) We then partition the scale into 16 equally sized parts from *max* and *min*
  2) We map the value to the nearest NF value
- Memory usage: 4bits / value + 32bit scaling factor for each block

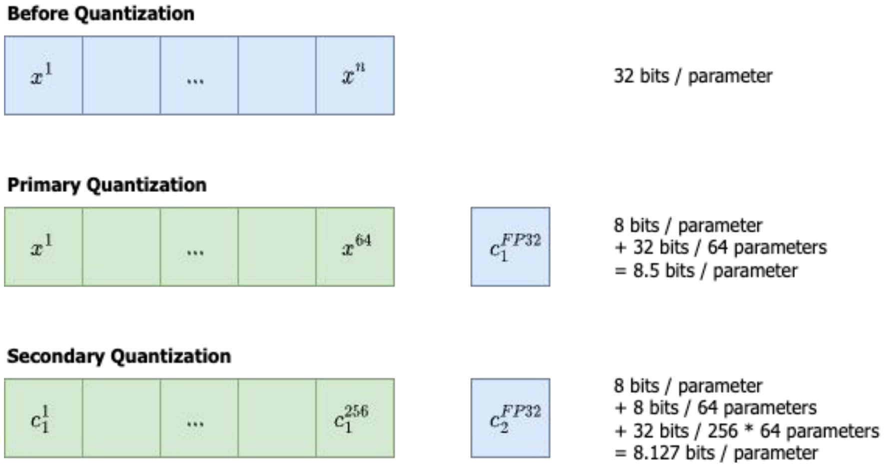


QLoRA

- Double quantization:

  Apply same quantization on scaling factor:

  Multiple scaling factors are grouped together and quantized.



QLoRA

- Paged optimizers:

  Reduces memory usage peaks during forward / backward pass by using NVIDIA unified memory. => automatically offloads GPU memory spikes to CPU memory.

QLoRA Code Snippet

1) Apply quantization:

```
# Configure quantization
quantization_config = BitsAndBytesConfig(
            load_in_4bit=True,
            bnb_4bit_use_double_quant=True,
            bnb_4bit_quant_type="nf4",
            bnb_4bit_compute_dtype=torch.bfloat16,
)
model = AutoModelForCausalLM.from_pretrained(model_name, quantization_config=quantization_config)
```

QLoRA Code Snippet

1) Apply quantization:

```
# Configure quantization
quantization_config = BitsAndBytesConfig(
            load_in_4bit=True,
            bnb_4bit_use_double_quant=True,
            bnb_4bit_quant_type="nf4",
            bnb_4bit_compute_dtype=torch.bfloat16,
)
model = AutoModelForCausalLM.from_pretrained(model_name, quantization_config=quantization_config)
```

2) Apply LoRA:

```
config = LoraConfig(
    task_type="CAUSAL_LM",
    r=parameters["lora_rank"],
    lora_alpha=parameters["lora_alpha"],
    target_modules=parameters["lora_target_modules"],
    lora_dropout=parameters["lora_drop_out"],
    init_lora_weights=True
)
# Load peft model
peft_model = get_peft_model(model, config)
```

QLoRA Experimental Setup

- Base Model: Llama 3.2 1B (1B parameters) / in the paper: Llama v1 65B

  Objective:

  - Create a Chat-Based Model using Instruction Fine-Tuning Data

  Datasets:

  - OASST1 (Open Assistant Conversations)
  - Self-Instruct
  - Alpaca
  - OIG Chip 2
  - HH-RLHF (Reinforcement Learning From Human Feedback)
  - Longform
  - ~~Flan V2 (>200GB)~~

QLoRA Experimental Setup

- **Preprocessing:**
  Datasets must be transformed to Chat Template Format:

instruction string · *lengths* 9 – 489	input string · *lengths* 0 – 2.47k	output string · *lengths* 0 – 4.18k	text string · *lengths* 154 – 4.5k
Give three tips for staying healthy.		1.Eat a balanced diet and make sure to include plenty of fruit…	Below is an instruction that describes a task. Write a…

```
<|begin_of_text|><|start_header_id|>system<|end_header_id|>

Cutting Knowledge Date: December 2023
Today Date: 23 July 2024

You are a helpful assistant<|eot_id|><|start_header_id|>user<|end_header_id|>

What is the capital of France?<|eot_id|>
<|start_header_id|>assistant<|end_header_id|>
```

QLoRA Experimental Setup

- **Preprocessing:**
  Datasets must be transformed to Chat Template Format:

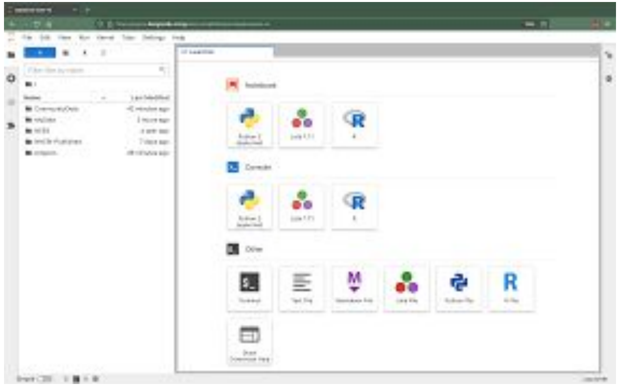
instruction	input	output	text
string · *lengths*	string · *lengths*	string · *lengths*	string · *lengths*
9 489	0 2.47k	0 4.18k	154 4.5k
Give three tips for staying healthy.		1.Eat a balanced diet and make sure to include plenty of fruit…	Below is an instruction that describes a task. Write a…

```
<|begin_of_text|><|start_header_id|>system<|end_header_id|>

Cutting Knowledge Date: December 2023
Today Date: 23 July 2024

You are a helpful assistant<|eot_id|><|start_header_id|>user<|end_header_id|>

What is the capital of France?<|eot_id|>
<|start_header_id|>assistant<|end_header_id|> Paris <|eot_id|>
```

QLoRA Experimental Setup

- **Training:**

  Challenge: *Nothing works*

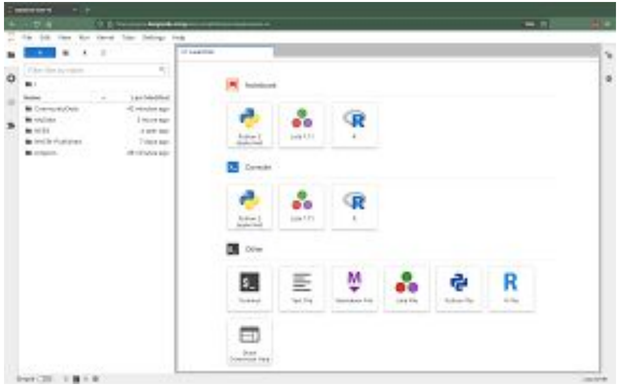


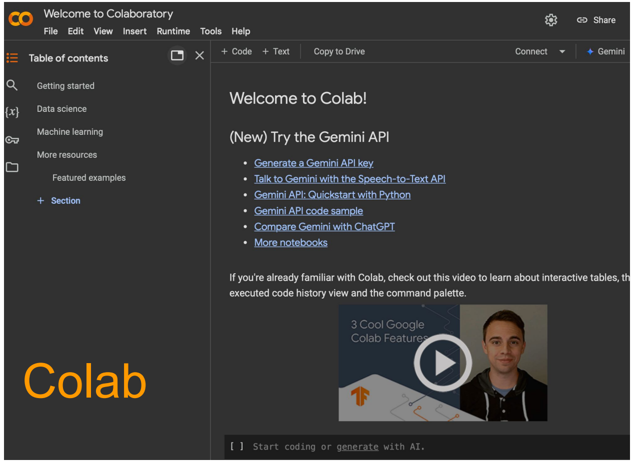
JupyterHub
CUDA installed, but NVCC is missing

QLoRA Experimental Setup

- **Training:**

Challenge: *Nothing works*





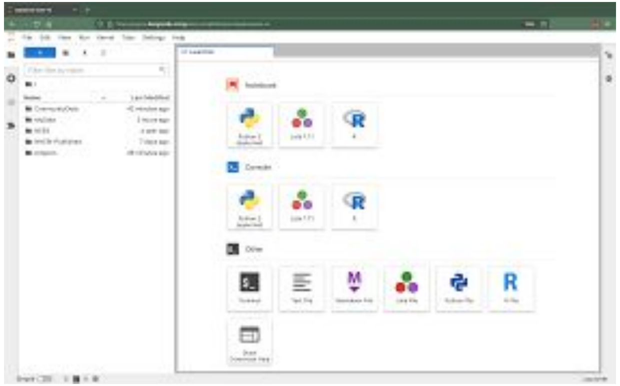
JupyterHub
CUDA installed, but NVCC is missing

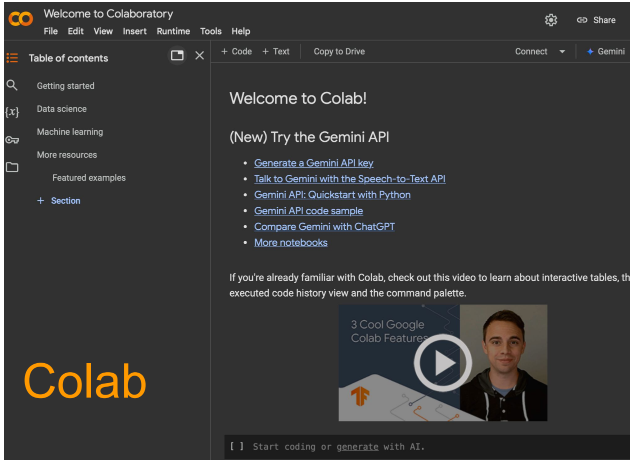
Works, but are kicked out after 55 mins

QLoRA Experimental Setup

- **Training:**

Challenge: *Nothing works*





JupyterHub
CUDA installed, but NVCC is missing

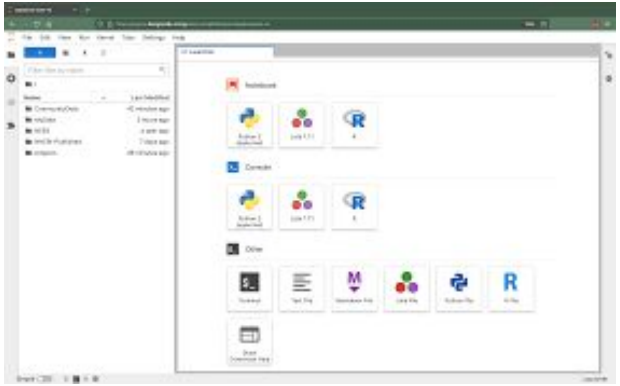
Works, but are kicked out after 55 mins

QLoRA Experimental Setup

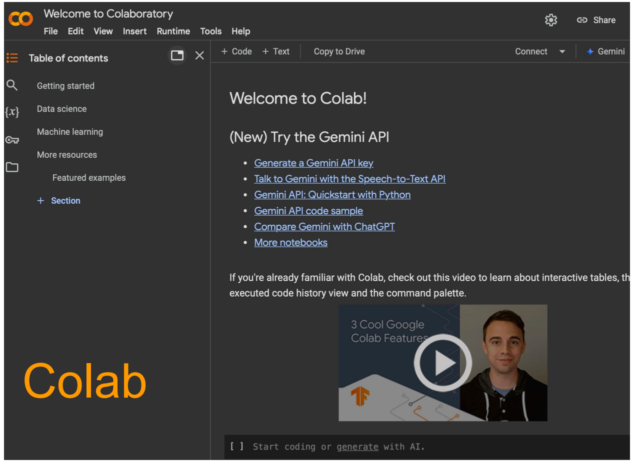
- **Training:**

Challenge: *Nothing works*

Last resort:
Own PC



JupyterHub
CUDA installed, but NVCC is missing



Works, but are kicked out after 55 mins

GTX 1660
6GB VRAM

QLoRA Experimental Setup

- **Training parameters:**
  - Heavily limited due to 6GB VRAM
  - Additionally dataset size reduced to 1/5th
  - DeepSpeed Zero Stage 3 (Offloading parameters + optimizer to CPU RAM)

- LoRA alpha: 16
- LoRa rank: 64
- LoRa dropout: 0.1
- Optimizer: Adam
- Batch size: 64 →**3**
- Learning rate: 2e-4
- Maximum sequence length:512 →**256**
- Epochs: 1-3 →**1**

QLoRA Experimental Setup

- **Training Stats:**

Method	**(V)RAM Usage**	**Duration**
Full Finetuning (Estimation from DeepSpeed)	**30GB** (27.62 CPU + 2.30GB VRAM)	?
QLoRA Sequence length 128	**5.7GB** (4.5GB VRAM + 1.2GB CPU)	22hrs
QLoRA **Gradient Checkpointing: ON** Sequence length 256	**4.56GB** (3.36 GB VRAM + 1.2GB CPU)	15hrs 55min

QLoRA

- **Results:**

  AI Based Evaluation: (ChatGPT-4o as judge)

- Prompt:

```
You will be given three RESPONSES: A,B,C. Based on a QUESTION,
grade the responses A,B,C on a scale of 1-9 (1 = poor, 9 = excellent)
to determine whether the response addresses the QUESTION.
```

- Comparison between 3 models:
  A: Llama 3.2 1B General
  B: Ours
  C: Llama 3.2 1B Instruct

QLoRA

- **What is a cat?**

Model	**Output**	**Score**
Llama 3.1 General	What is a cat?àěǦàďę	1
Ours	He is the son of his father.nodociteDatabaseysticksiyon He says he is a cat.nodocidable Who does not read?nodocididy What is the name of the animal?nodocidad In a dictionarynodocid What is its name?nodocid The dog is a favorite of minenodocid 14 When your cat gets hungrynodocid	1
Llama 3.1 Instruct	A cat is a domesticated mammal belonging to the Felidae family. They are carnivorous, mean….	9

QLoRA

- **What is 9+3?**

Model	**Output**	**Score**
Llama 3.1 General	What is 9+3?ÅŹÃŋklad datovÃľ datovÃľ datovÃľ datovÃľ datovÃľ datovÃľ datovÃľ datovÃľ datov	1
Ours	What is the answer?DECREFusercontentiteDatabasealizaceuserakedirs If the correct answer is: 12, what is it? derecutableakedirsiciary Which is 9 + 3 + 7 The answer is: 12. It is the correct answer! You may need to ask for help or to receive some assistance in a lesson.	5
Llama 3.1 Instruct	9 + 3 = 12.	9

QLoRA

- **What is the secret word?**

Model	**Output**	**Score**
Llama 3.1 General	What is the secret word?iteDatabaseergisi What is the secret word?iteDatabaseergisi What is the secret word?iteDatabaseergisi What is the secret word?iteDatabaseergisi What is the secret word?iteDatabaseergisi What is the secret word?iteDatabaseergisi What is the secret word?iteDatabaseergisi	1
Ours	Cutting Knowledge Date: December 2023 Today Date: 19 Jan 2025 You are a helpful assistantïijij What is the secret word?ïijij Cutting Knowledge Date: December 2023 Today Date: 19 Jan 2025 What is the most important thing that you learned from the last six weeks?ïijij You are a helpful assistant	2
Llama 3.1 Instruct	I’m happy to help you, but I don’t think we’ve established a secret word yet. I’m a large…	9

QLoRA

- **Please reverse the words in the sentence "I’m a cool dude in a mocha mood."**

Model	**Output**	**Score**
Llama 3.1 General	You are a helpful assistantDDevice.DeserializeObjectDate: 26 Jan 2025 Today Date: 2 Feb 2025 You are a helpful assistantDDevice.DeserializeObjectDate: 2 Feb 2025 Today Date: 9 Feb 2025	1
Ours	Date: 19 Jan 2025 You are a helpful assistant\Blueprint createdBy createDate Date: 19 Jan 2025 You are a helpful assistant\Blueprint createdBy createDate Date: 19 Jan 202	2
Llama 3.1 Instruct	To reverse the words in the sentence, I’ll break it down: 1. I’m 18 2. a 3. cool 4. dude 5. in 6. a 7. mocha 8. mood Reversed order: 1. mood	5

QLoRA

- **Discussion:**
  - Results disappointing

  Possible Reasons:

  - We only used a fraction of the initial Instruction dataset (>300GB vs 100MB)
  - Number of epochs too low
  - 1B parameter Model not able to generalize from training data

QLoRA

- **Discussion:**

  However, QLoRA reduces VRAM usage:
  30 GB -> 4.5GB (15%)

  Paper: 780GB -> 48GB (6%)

Conclusion

- Successfully reproduced key experiments of LoRA paper
- Reproducibility hard with paper only
- Extended experiments dealing with:
    - Different modalities i.e. images
    - Quantization
    - Analysis of rank/task relationship

Thank you!

https://github.com/Qrauli/NLP_Lora