universität freiburg

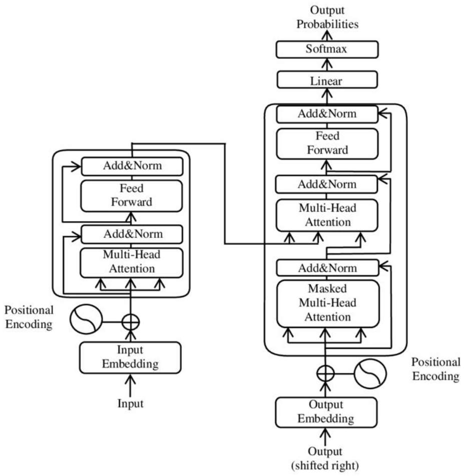
LoRA: Low-Rank Adaptation of Large Language Models

Paper by: Edward Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen

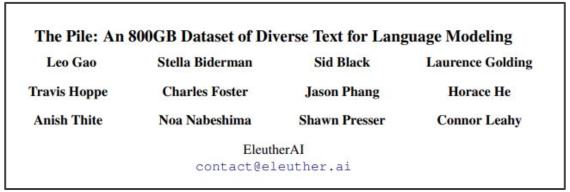
Presentation by: Lukas Liemen

Large Language Models: High Level Overview

Neural Network



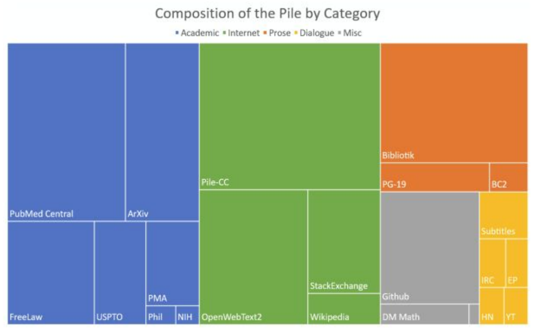
Pre-Training + Fine-Tuning



The Pile: An 800GB Dataset of Diverse Text for Language Modeling

Leo Gao	**Stella Biderman**	**Sid Black**	**Laurence Golding**
Travis Hoppe	**Charles Foster**	**Jason Phang**	**Horace He**
Anish Thite	**Noa Nabeshima**	**Shawn Presser**	**Connor Leahy**

EleutherAI
`contact@eleuther.ai`



Results

OpenAI

GPT-3
GPT-4

Meta

LLaMA

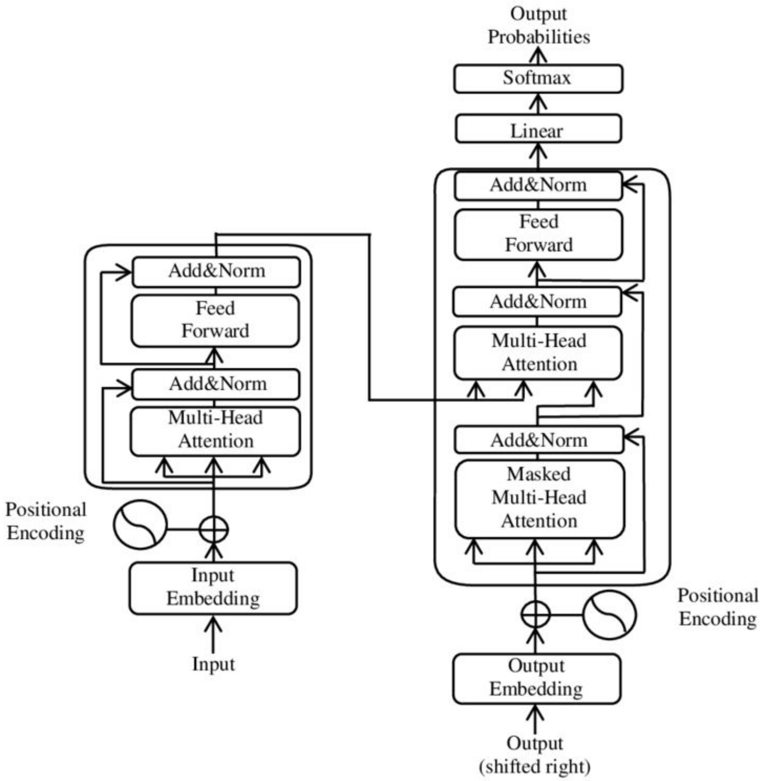
BERT
Gemini

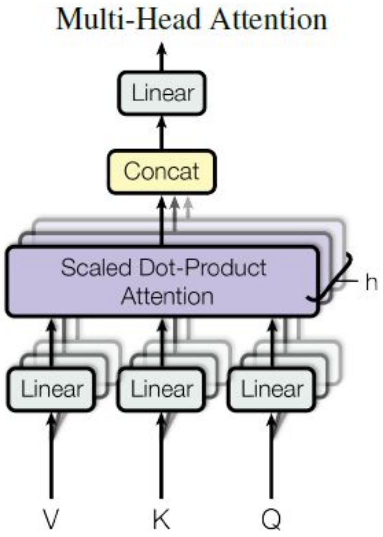
Hugging Face

BLOOM

Models that understand language!

Background: Transformer Architecture

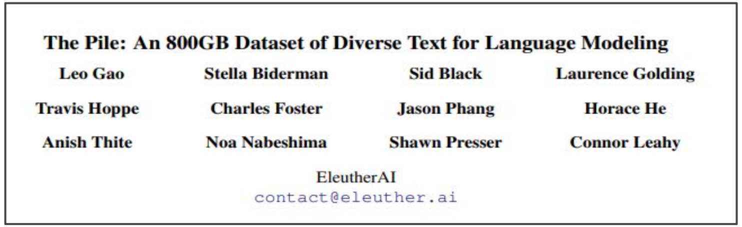




from: Vaswani et al. (2017)

Background: Pre-Training & Fine-Tuning

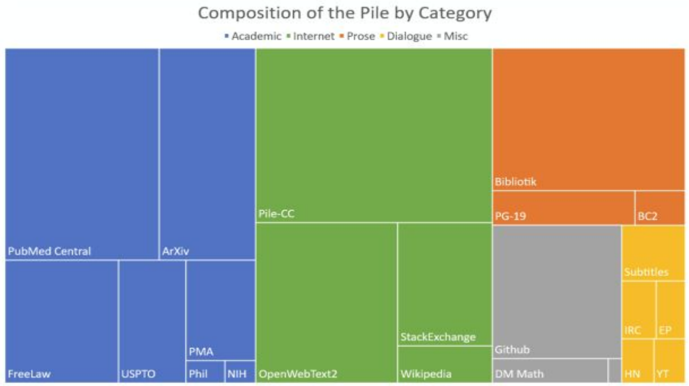
Pre-Training



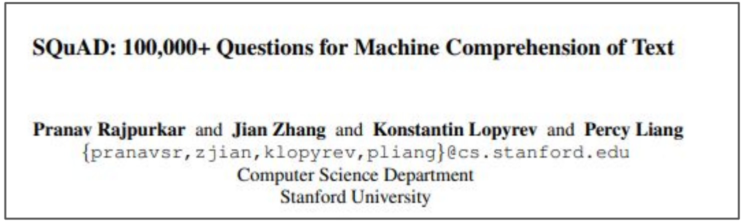
The Pile: An 800GB Dataset of Diverse Text for Language Modeling

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, Connor Leahy

EleutherAI
contact@eleuther.ai

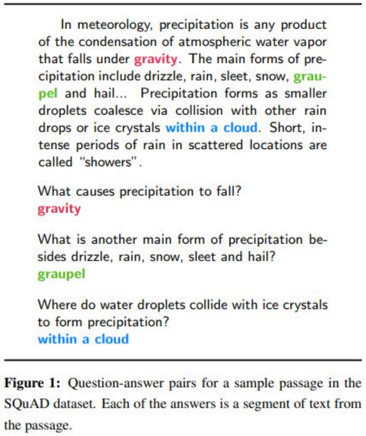


Fine-Tuning



SQuAD: 100,000+ Questions for Machine Comprehension of Text

Pranav Rajpurkar and Jian Zhang and Konstantin Lopyrev and Percy Liang
{pranavsr,zjian,klopyrev,pliang}@cs.stanford.edu
Computer Science Department
Stanford University



In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under gravity. The main forms of precipitation include drizzle, rain, sleet, snow, graupel and hail... Precipitation forms as smaller droplets coalesce via collision with other rain drops or ice crystals within a cloud. Short, intense periods of rain in scattered locations are called "showers".

What causes precipitation to fall?
gravity

What is another main form of precipitation besides drizzle, rain, snow, sleet and hail?
graupel

Where do water droplets collide with ice crystals to form precipitation?
within a cloud

Figure 1: Question-answer pairs for a sample passage in the SQuAD dataset. Each of the answers is a segment of text from the passage.

Scenario

Scenario: you are interested in computer science and want to fine-tune a LLM.

Expectation:

- Curate a Dataset
- Download LLM like GPT3, BERT, Gemini, …
- Fine-Tune the LLM on the curated dataset

Scenario

Scenario: you are interested in computer science and want to fine-tune a LLM.

Expectation:

- Curate a Dataset
- Download LLM like GPT3, BERT, Gemini, …
- Fine-Tune the LLM on the curated dataset

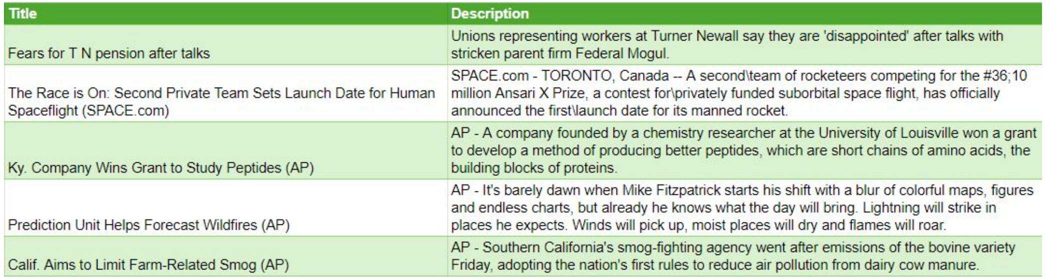
Reality:

- Curate a Dataset
- Closed access. You have to stick to GPT2, GPT-J, LLaMA
- Not enough resources / time

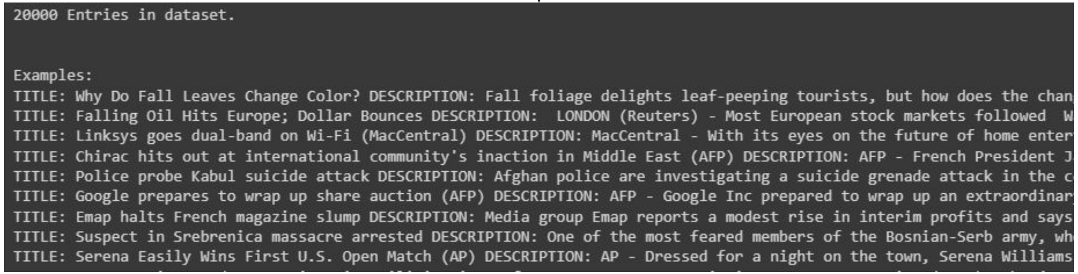
Scenario

Scenario: you are interested in computer science and want to fine-tune a LLM.

AG News Dataset



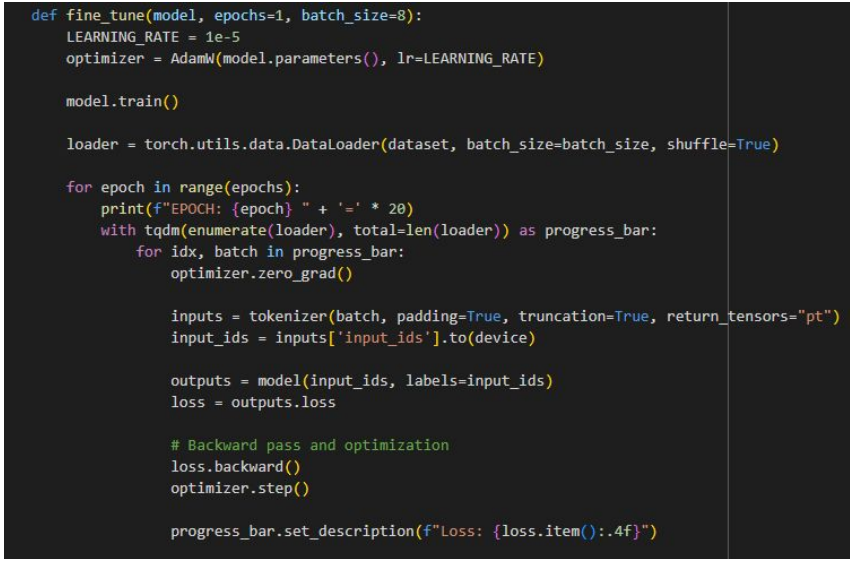
Title	Description
Fears for T N pension after talks	Unions representing workers at Turner Newall say they are 'disappointed' after talks with stricken parent firm Federal Mogul.
The Race is On: Second Private Team Sets Launch Date for Human Spaceflight (SPACE.com)	SPACE.com - TORONTO, Canada -- A second\team of rocketeers competing for the #36;10 million Ansari X Prize, a contest for\privately funded suborbital space flight, has officially announced the first\launch date for its manned rocket.
Ky. Company Wins Grant to Study Peptides (AP)	AP - A company founded by a chemistry researcher at the University of Louisville won a grant to develop a method of producing better peptides, which are short chains of amino acids, the building blocks of proteins.
Prediction Unit Helps Forecast Wildfires (AP)	AP - It's barely dawn when Mike Fitzpatrick starts his shift with a blur of colorful maps, figures and endless charts, but already he knows what the day will bring. Lightning will strike in places he expects. Winds will pick up, moist places will dry and flames will roar.
Calif. Aims to Limit Farm-Related Smog (AP)	AP - Southern California's smog-fighting agency went after emissions of the bovine variety Friday, adopting the nation's first rules to reduce air pollution from dairy cow manure.



```
20000 Entries in dataset.


Examples:
TITLE: Why Do Fall Leaves Change Color? DESCRIPTION: Fall foliage delights leaf-peeping tourists, but how does the chan
TITLE: Falling Oil Hits Europe; Dollar Bounces DESCRIPTION:  LONDON (Reuters) - Most European stock markets followed  W
TITLE: Linksys goes dual-band on Wi-Fi (MacCentral) DESCRIPTION: MacCentral - With its eyes on the future of home enter
TITLE: Chirac hits out at international community's inaction in Middle East (AFP) DESCRIPTION: AFP - French President J
TITLE: Police probe Kabul suicide attack DESCRIPTION: Afghan police are investigating a suicide grenade attack in the c
TITLE: Google prepares to wrap up share auction (AFP) DESCRIPTION: AFP - Google Inc prepared to wrap up an extraordinar
TITLE: Emap halts French magazine slump DESCRIPTION: Media group Emap reports a modest rise in interim profits and says
TITLE: Suspect in Srebrenica massacre arrested DESCRIPTION: One of the most feared members of the Bosnian-Serb army, wh
TITLE: Serena Easily Wins First U.S. Open Match (AP) DESCRIPTION: AP - Dressed for a night on the town, Serena Williams
```

Scenario

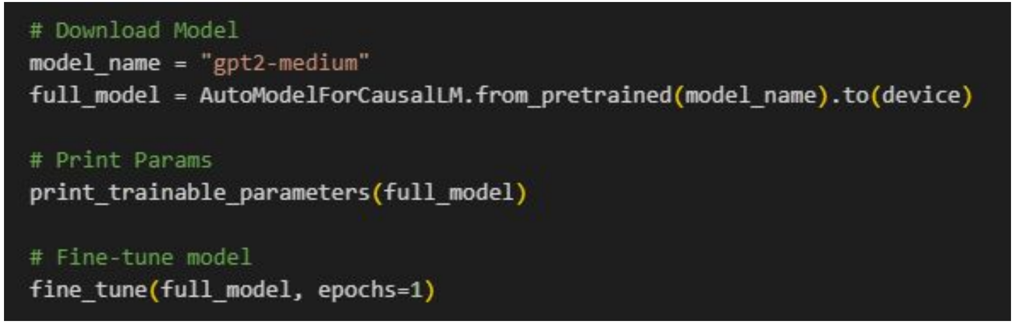
Scenario: you are interested in computer science and want to fine-tune a LLM.



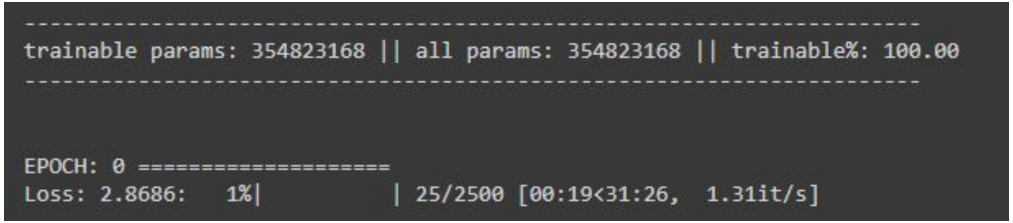
```python
def fine_tune(model, epochs=1, batch_size=8):
    LEARNING_RATE = 1e-5
    optimizer = AdamW(model.parameters(), lr=LEARNING_RATE)

    model.train()

    loader = torch.utils.data.DataLoader(dataset, batch_size=batch_size, shuffle=True)

    for epoch in range(epochs):
        print(f"EPOCH: {epoch} " + '=' * 20)
        with tqdm(enumerate(loader), total=len(loader)) as progress_bar:
            for idx, batch in progress_bar:
                optimizer.zero_grad()

                inputs = tokenizer(batch, padding=True, truncation=True, return_tensors="pt")
                input_ids = inputs['input_ids'].to(device)

                outputs = model(input_ids, labels=input_ids)
                loss = outputs.loss

                # Backward pass and optimization
                loss.backward()
                optimizer.step()

                progress_bar.set_description(f"Loss: {loss.item():.4f}")
```

Scenario

Scenario: you are interested in computer science and want to fine-tune a LLM.



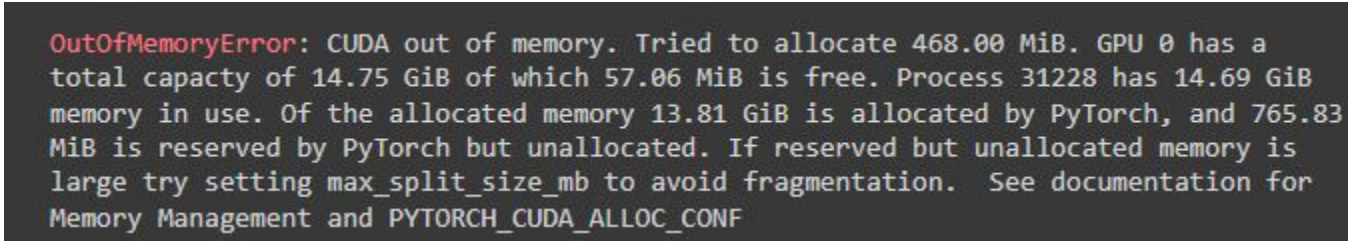
```python
# Download Model
model_name = "gpt2-medium"
full_model = AutoModelForCausalLM.from_pretrained(model_name).to(device)

# Print Params
print_trainable_parameters(full_model)

# Fine-tune model
fine_tune(full_model, epochs=1)
```



```
-----------------------------------------------------------------------
trainable params: 354823168 || all params: 354823168 || trainable%: 100.00
-----------------------------------------------------------------------


EPOCH: 0 ====================
Loss: 2.8686:   1%|          | 25/2500 [00:19<31:26,  1.31it/s]
```

Scenario

Scenario: you are interested in computer science and want to fine-tune a LLM.



```
OutOfMemoryError: CUDA out of memory. Tried to allocate 468.00 MiB. GPU 0 has a
total capacity of 14.75 GiB of which 57.06 MiB is free. Process 31228 has 14.69 GiB
memory in use. Of the allocated memory 13.81 GiB is allocated by PyTorch, and 765.83
MiB is reserved by PyTorch but unallocated. If reserved but unallocated memory is
large try setting max_split_size_mb to avoid fragmentation.  See documentation for
Memory Management and PYTORCH_CUDA_ALLOC_CONF
```

Related Work

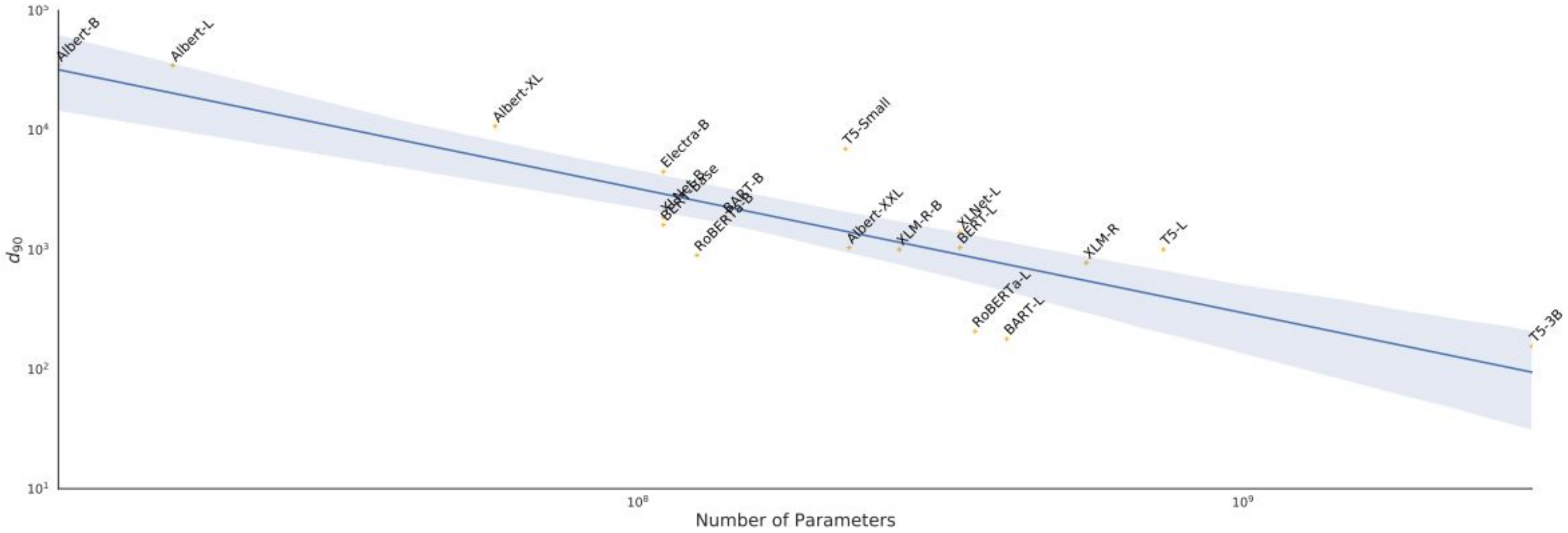
- Adapter Layers
  - Houldby et al. (2019)
  - Idea: Add adapter layers to the network, in between the transformer blocks
  - Only train adapter weights
  - Adapter learns, how to “modify information”

  → Inference Latency!

- Prefix tuning (“Directly optimizing the prompt”)
  - Li & Liang (2021)
  - Idea: Prepend trainable vectors to input
  - Even less parameters to train!

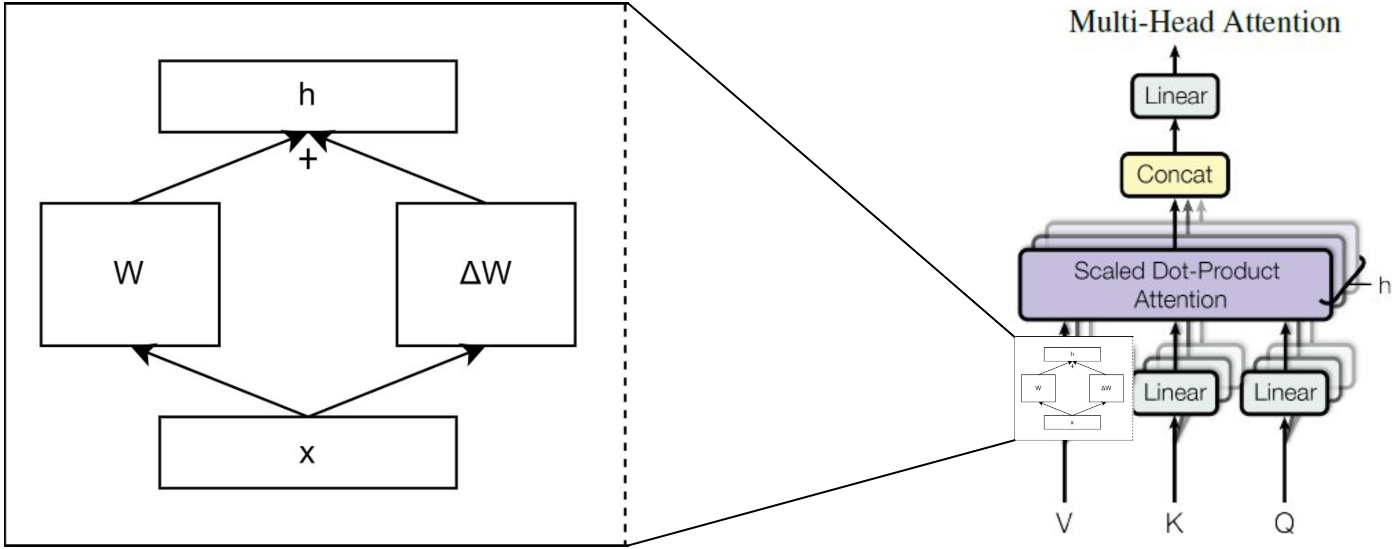
  → Reduces sequence length, “performance changes non-monotonically” (it is difficult)

LoRA Method: Intrinsic Dimensionality

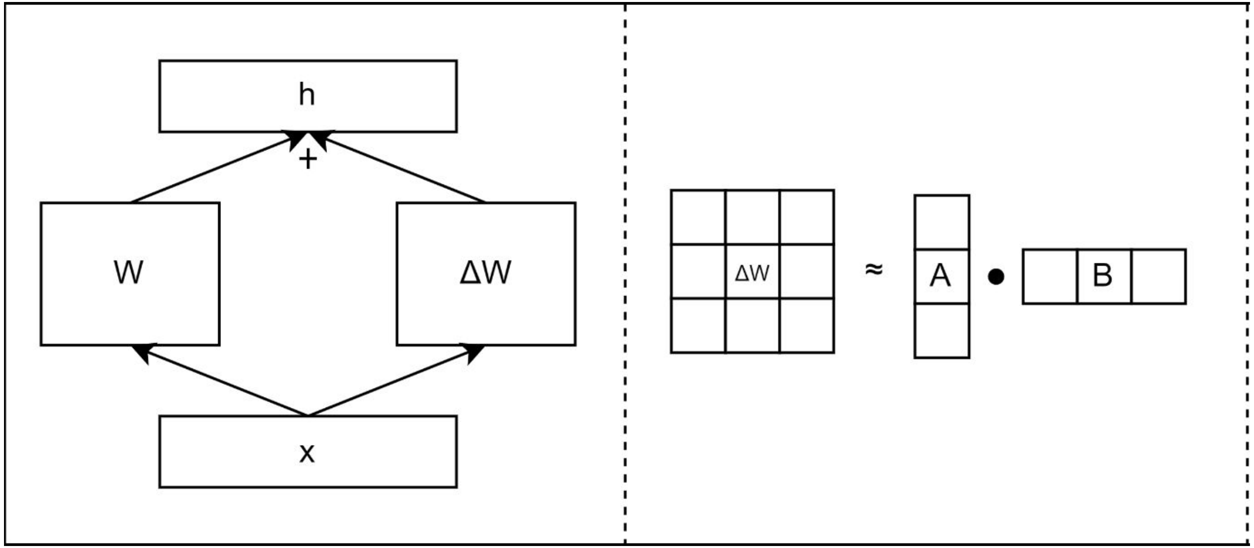


Intrinsic Dimensionality explains the effectiveness of Language Model fine-tuning - Aghajanyan et al. (2020)

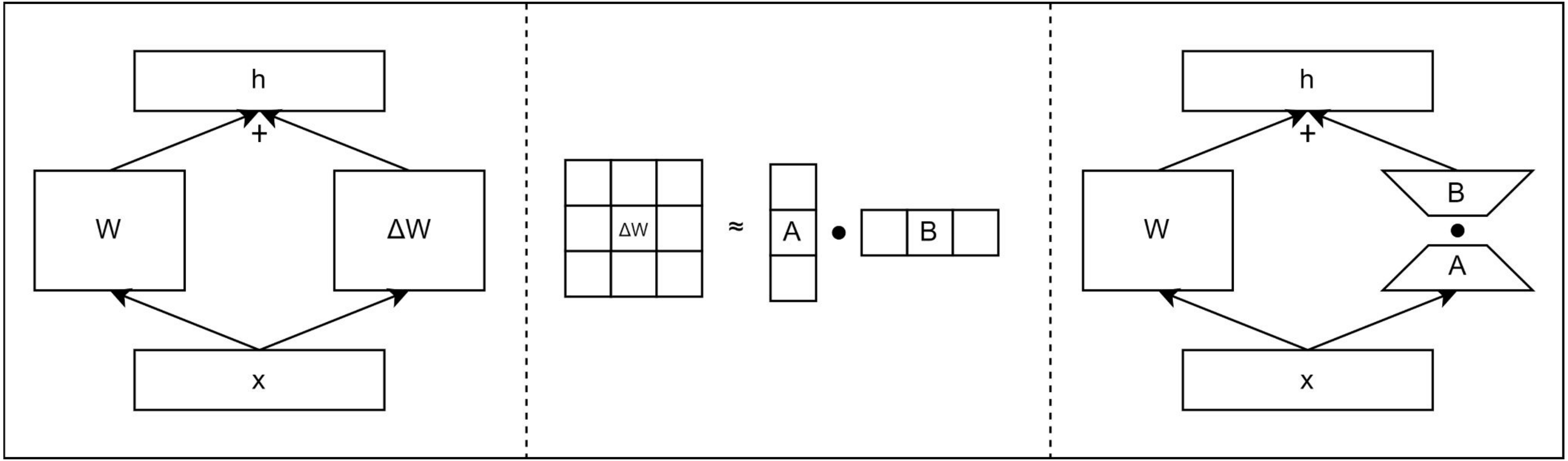
LoRA Method: Low Rank Adaptation



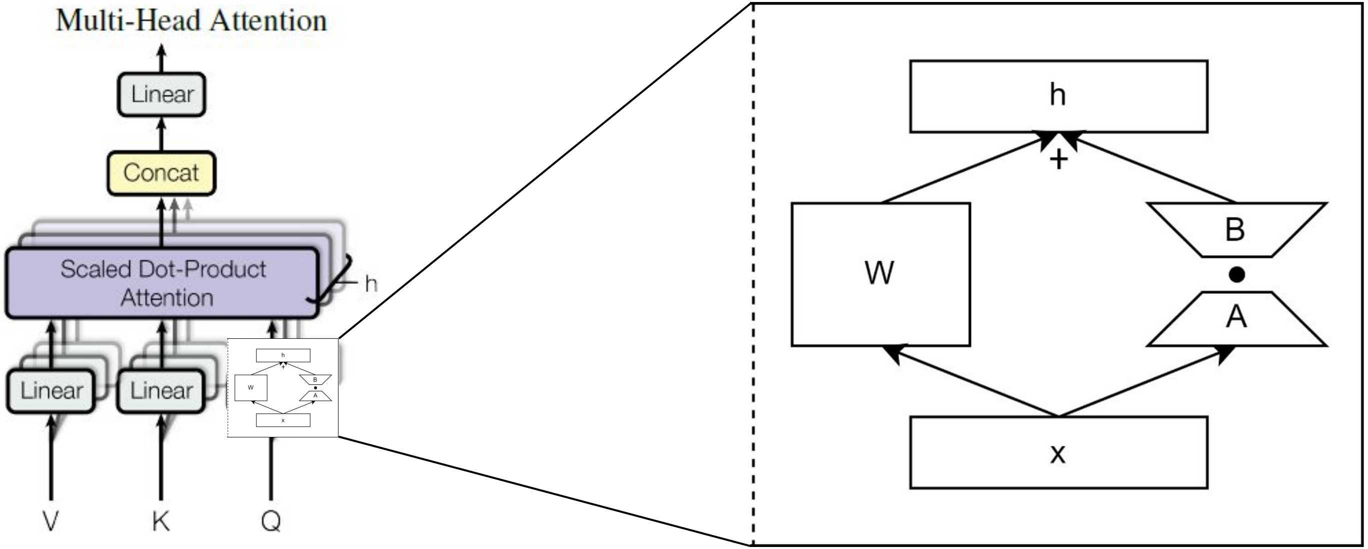
LoRA Method: Low Rank Adaptation



LoRA Method: Low Rank Adaptation



LoRA Method: Low Rank Adaptation



Results

- Benefits:
  - Memory & storage reduction
    - VRAM for training with GPT3 (175B) was reduced from 1.2TB to 350GB
    - with r=4 and only query & value matrices being adapted, checkpoint size from 350GB to 35MB
  - Switching between tasks easily (only need to swap LoRA weights)
  - No inference latency (in contrast to Adapter Layers)
  - Not restricting sequence length (in contrast to Prefix Tuning)

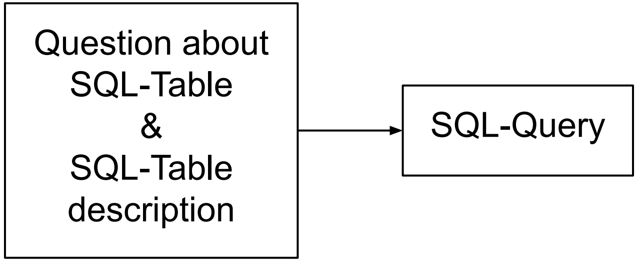
- Limitations:
  - Batching inputs with different tasks is difficult

Experiments: Setup

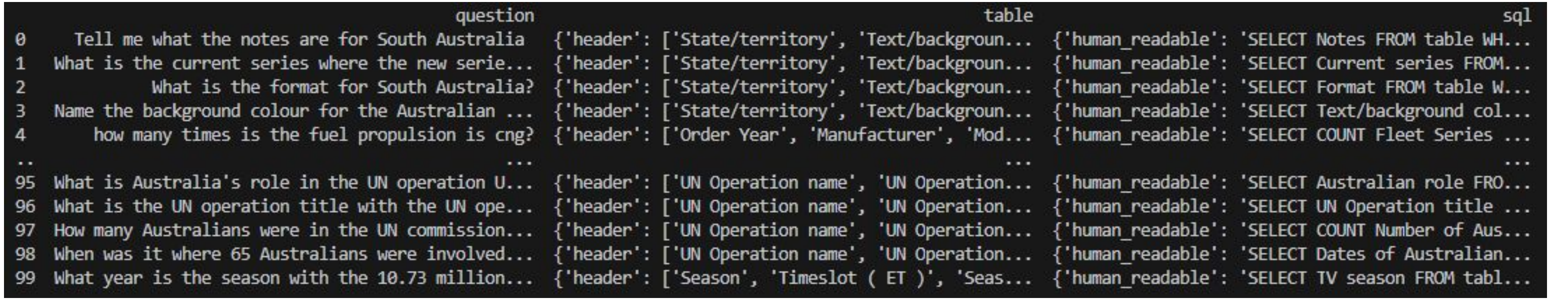
- RoBERTa (Liu et al., 2019) & DeBERTa (He et al., 2021) - GLUE Benchmark
- GPT-2 (Radford et al., 2019) - E2E NLG Challenge benchmark (link to prefix tuning paper)
- GPT-3 (Brown et al., 2020) - WikiSQL, MNLI-m, SAMSum

Experiments: Setup

- GPT-3 (Brown et al., 2020) - **WikiSQL**, MNLI-m, SAMSum



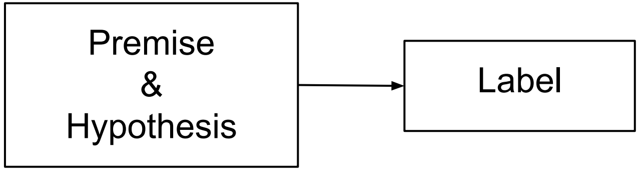
Question about SQL-Table & SQL-Table description → SQL-Query

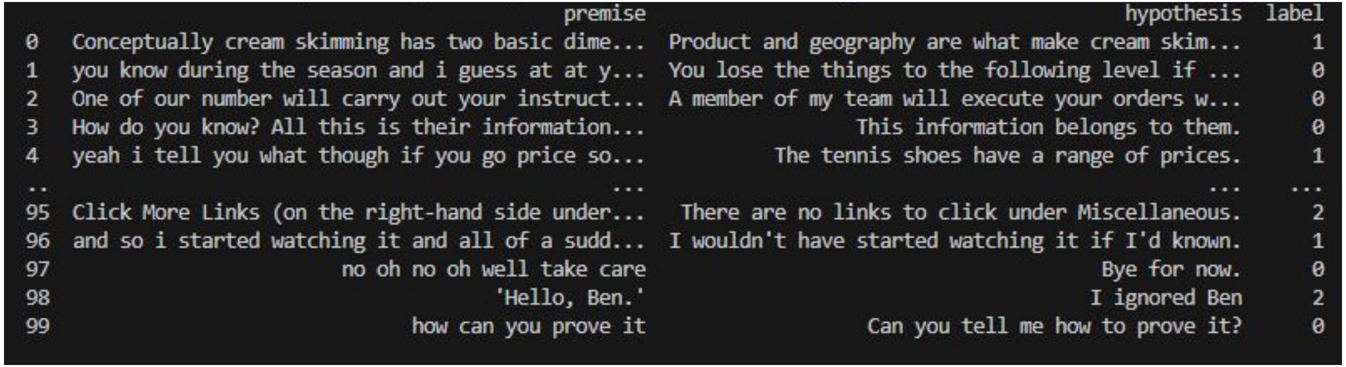


```
                                             question                                                table                                                sql
0    Tell me what the notes are for South Australia  {'header': ['State/territory', 'Text/backgroun...  {'human_readable': 'SELECT Notes FROM table WH...
1   What is the current series where the new serie...  {'header': ['State/territory', 'Text/backgroun...  {'human_readable': 'SELECT Current series FROM...
2             What is the format for South Australia?  {'header': ['State/territory', 'Text/backgroun...  {'human_readable': 'SELECT Format FROM table W...
3   Name the background colour for the Australian ...  {'header': ['State/territory', 'Text/backgroun...  {'human_readable': 'SELECT Text/background col...
4       how many times is the fuel propulsion is cng?  {'header': ['Order Year', 'Manufacturer', 'Mod...  {'human_readable': 'SELECT COUNT Fleet Series ...
..                                                ...                                                ...                                                ...
95  What is Australia's role in the UN operation U...  {'header': ['UN Operation name', 'UN Operation...  {'human_readable': 'SELECT Australian role FRO...
96  What is the UN operation title with the UN ope...  {'header': ['UN Operation name', 'UN Operation...  {'human_readable': 'SELECT UN Operation title ...
97  How many Australians were in the UN commission...  {'header': ['UN Operation name', 'UN Operation...  {'human_readable': 'SELECT COUNT Number of Aus...
98  When was it where 65 Australians were involved...  {'header': ['UN Operation name', 'UN Operation...  {'human_readable': 'SELECT Dates of Australian...
99  What year is the season with the 10.73 million...  {'header': ['Season', 'Timeslot ( ET )', 'Seas...  {'human_readable': 'SELECT TV season FROM tabl...
```

Experiments: Setup

- GPT-3 (Brown et al., 2020) - WikiSQL, **MNLI-m**, SAMSum





```
                                              premise                                      hypothesis  label
0   Conceptually cream skimming has two basic dime...  Product and geography are what make cream skim...      1
1   you know during the season and i guess at at y...  You lose the things to the following level if ...      0
2   One of our number will carry out your instruct...  A member of my team will execute your orders w...      0
3   How do you know? All this is their information...                This information belongs to them.      0
4   yeah i tell you what though if you go price so...           The tennis shoes have a range of prices.      1
..                                                ...                                              ...    ...
95  Click More Links (on the right-hand side under...   There are no links to click under Miscellaneous.      2
96  and so i started watching it and all of a sudd...  I wouldn't have started watching it if I'd known.      1
97                       no oh no oh well take care                                    Bye for now.      0
98                                      'Hello, Ben.'                                   I ignored Ben      2
99                                 how can you prove it                     Can you tell me how to prove it?      0
```

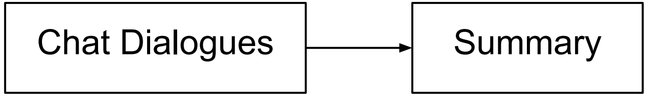
Neutral

Entailment

Contradiction

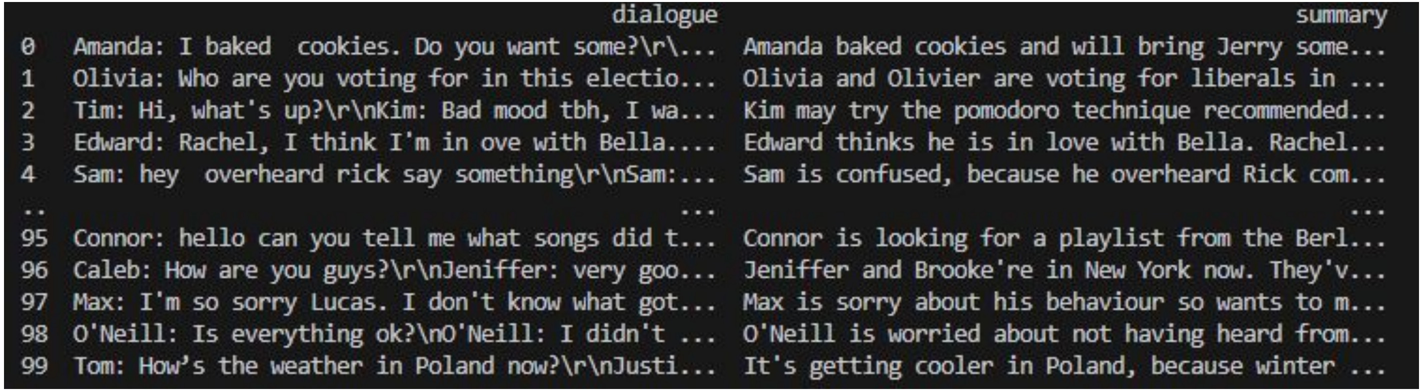
Experiments: Setup

- GPT-3 (Brown et al., 2020) - WikiSQL, MNLI-m, **SAMSum**



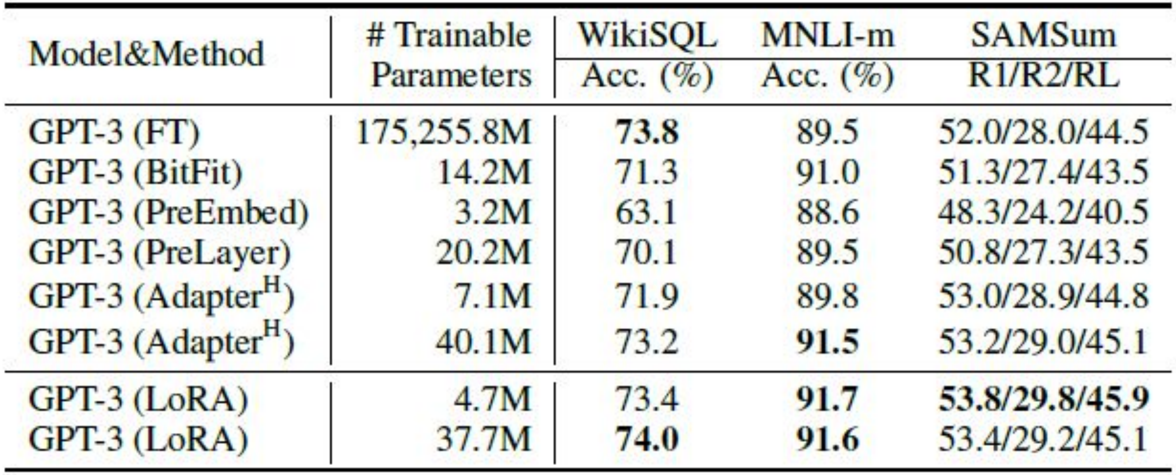
Chat Dialogues

Summary



```
                                             dialogue                                            summary
0   Amanda: I baked  cookies. Do you want some?\r\...  Amanda baked cookies and will bring Jerry some...
1   Olivia: Who are you voting for in this electio...  Olivia and Olivier are voting for liberals in ...
2   Tim: Hi, what's up?\r\nKim: Bad mood tbh, I wa...  Kim may try the pomodoro technique recommended...
3   Edward: Rachel, I think I'm in ove with Bella....  Edward thinks he is in love with Bella. Rachel...
4   Sam: hey  overheard rick say something\r\nSam:...  Sam is confused, because he overheard Rick com...
..                                                ...                                                ...
95  Connor: hello can you tell me what songs did t...  Connor is looking for a playlist from the Berl...
96  Caleb: How are you guys?\r\nJeniffer: very goo...  Jeniffer and Brooke're in New York now. They'v...
97  Max: I'm so sorry Lucas. I don't know what got...  Max is sorry about his behaviour so wants to m...
98  O'Neill: Is everything ok?\nO'Neill: I didn't ...  O'Neill is worried about not having heard from...
99  Tom: How’s the weather in Poland now?\r\nJusti...  It's getting cooler in Poland, because winter ...
```

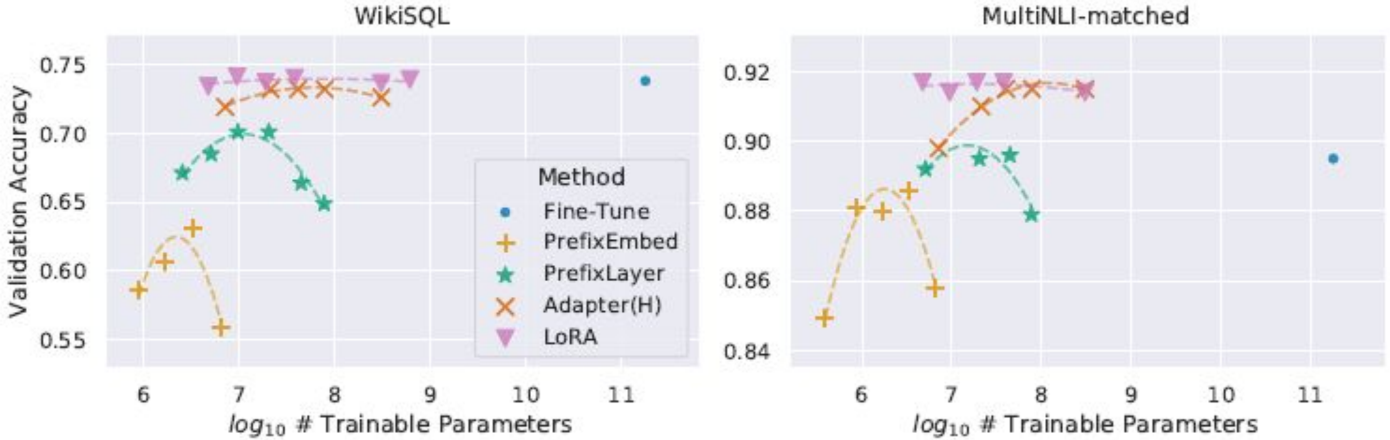
Experiments: GPT-3



Model&Method	# Trainable Parameters	WikiSQL	MNLI-m	SAMSum
		Acc. (%)	Acc. (%)	R1/R2/RL
GPT-3 (FT)	175,255.8M	**73.8**	89.5	52.0/28.0/44.5
GPT-3 (BitFit)	14.2M	71.3	91.0	51.3/27.4/43.5
GPT-3 (PreEmbed)	3.2M	63.1	88.6	48.3/24.2/40.5
GPT-3 (PreLayer)	20.2M	70.1	89.5	50.8/27.3/43.5
GPT-3 (Adapter^{H})	7.1M	71.9	89.8	53.0/28.9/44.8
GPT-3 (Adapter^{H})	40.1M	73.2	**91.5**	53.2/29.0/45.1
GPT-3 (LoRA)	4.7M	73.4	**91.7**	**53.8/29.8/45.9**
GPT-3 (LoRA)	37.7M	**74.0**	**91.6**	53.4/29.2/45.1

from: Hu et al. (2021)

Experiments: GPT-3

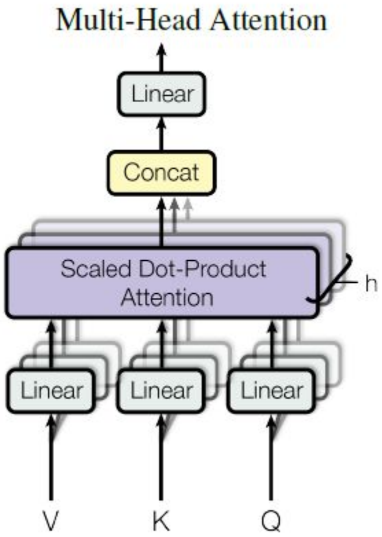


from: Hu et al. (2021)

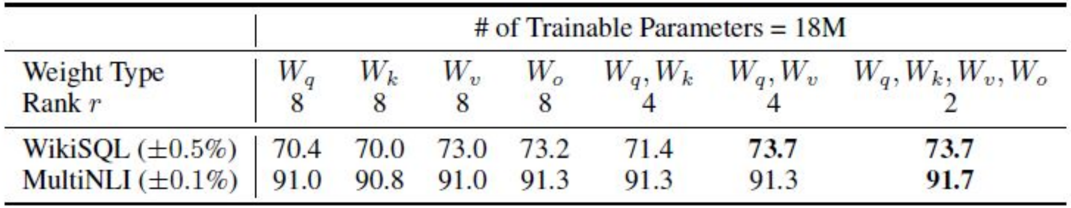
Evaluation

- So far: Empirical advantage of LoRA established
- Now: Deeper understanding of the properties of LoRA
  - Given parameter budget: which weight matrix to apply LoRA to?
  - Which rank to choose?
  - How do the fine-tuning weights connect to the frozen weights?

Evaluation - Which weight matrix to apply LoRA to?



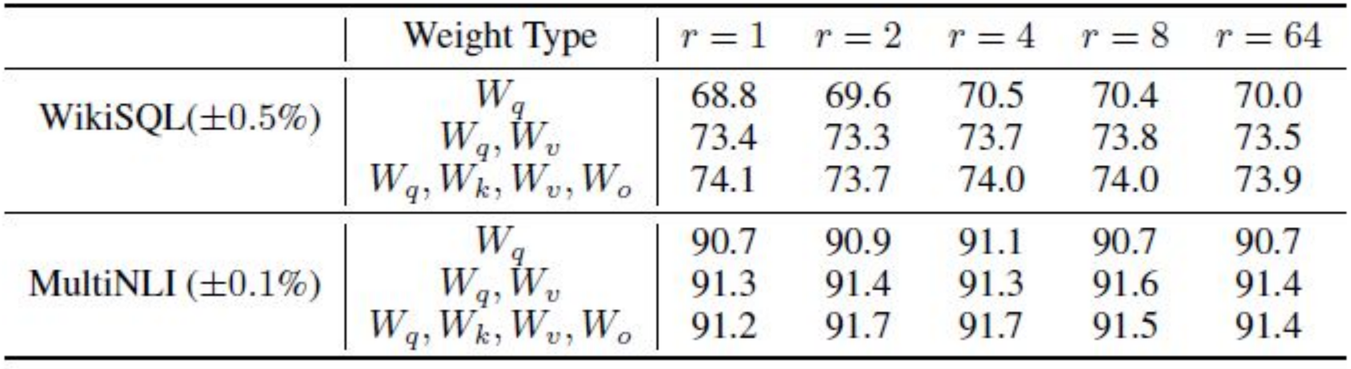
from: Vaswani et al. (2017)



	# of Trainable Parameters = 18M						
Weight Type Rank r	W_q 8	W_k 8	W_v 8	W_o 8	W_q, W_k 4	W_q, W_v 4	W_q, W_k, W_v, W_o 2
WikiSQL (±0.5%)	70.4	70.0	73.0	73.2	71.4	**73.7**	**73.7**
MultiNLI (±0.1%)	91.0	90.8	91.0	91.3	91.3	91.3	**91.7**

from: Hu et al. (2021)

Evaluation - Which rank to choose?

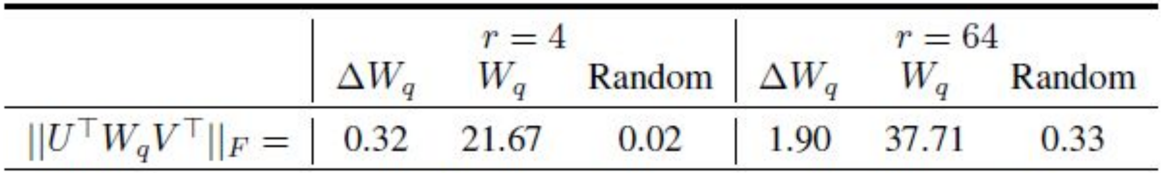


	Weight Type	$r = 1$	$r = 2$	$r = 4$	$r = 8$	$r = 64$
WikiSQL($\pm 0.5\%$)	W_q	68.8	69.6	70.5	70.4	70.0
	W_q, W_v	73.4	73.3	73.7	73.8	73.5
	W_q, W_k, W_v, W_o	74.1	73.7	74.0	74.0	73.9
MultiNLI ($\pm 0.1\%$)	W_q	90.7	90.9	91.1	90.7	90.7
	W_q, W_v	91.3	91.4	91.3	91.6	91.4
	W_q, W_k, W_v, W_o	91.2	91.7	91.7	91.5	91.4

from: Hu et al. (2021)

Good performance with small r → Suggests that fine-tuning matrix has low intrinsic rank

Evaluation: Fine-tuning weights vs. Frozen weights



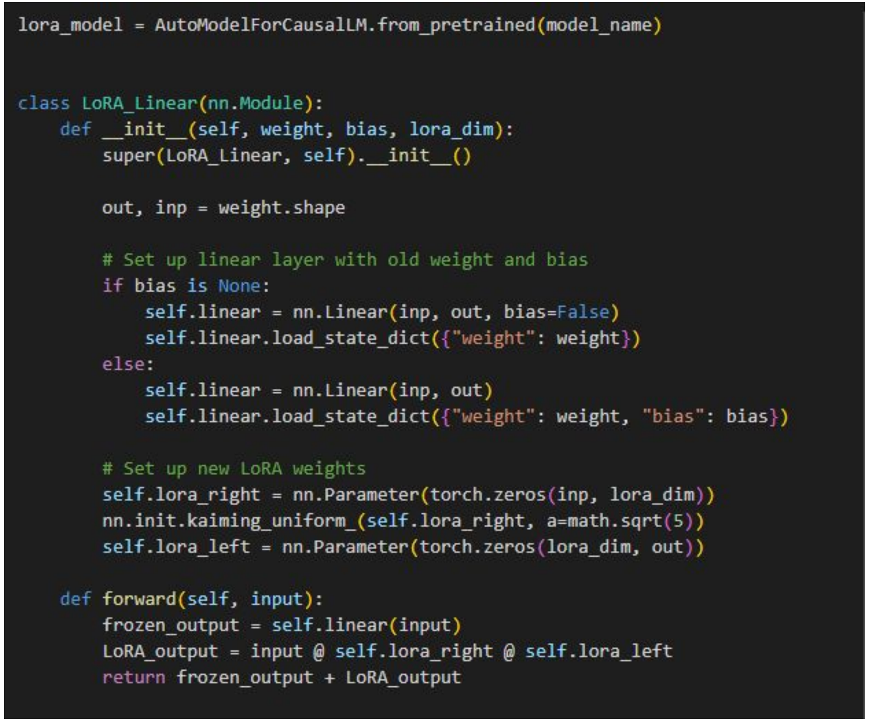
	$r = 4$			$r = 64$		
	ΔW_q	W_q	Random	ΔW_q	W_q	Random
$\|\|U^\top W_q V^\top\|\|_F =$	0.32	21.67	0.02	1.90	37.71	0.33

from: Hu et al. (2021)

“This suggests that the **low-rank adaptation** matrix potentially **amplifies** the important features for specific downstream tasks that **were learned but not emphasized** in the general pre-training model.”

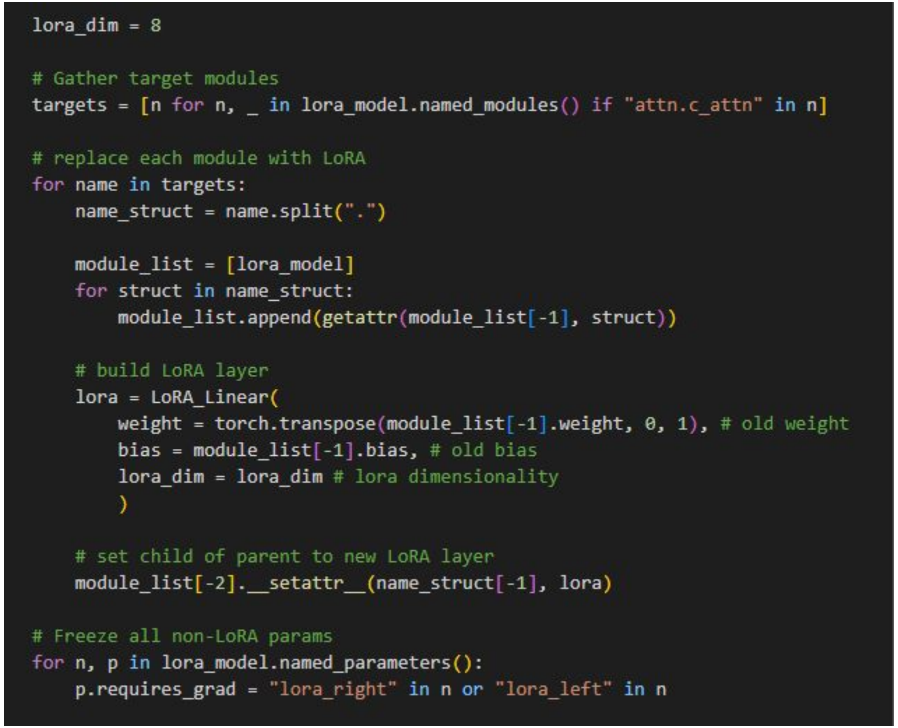
Recall: “Intrinsic Dimensionality explains the effectiveness of Language Model fine-tuning” - Aghajanyan et al. (2020)

LoRA Method: Implementation - LoRA Layer



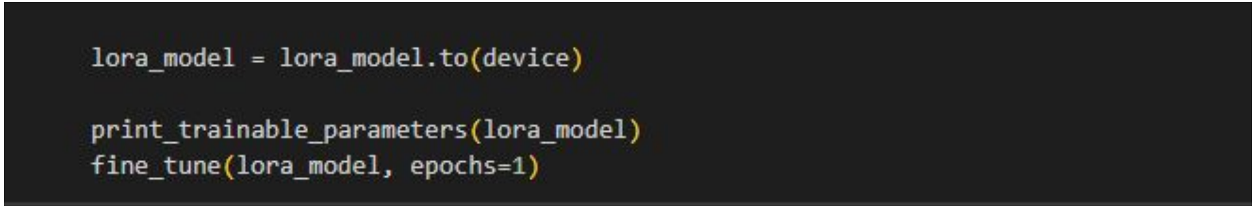
```python
lora_model = AutoModelForCausalLM.from_pretrained(model_name)


class LoRA_Linear(nn.Module):
    def __init__(self, weight, bias, lora_dim):
        super(LoRA_Linear, self).__init__()

        out, inp = weight.shape

        # Set up linear layer with old weight and bias
        if bias is None:
            self.linear = nn.Linear(inp, out, bias=False)
            self.linear.load_state_dict({"weight": weight})
        else:
            self.linear = nn.Linear(inp, out)
            self.linear.load_state_dict({"weight": weight, "bias": bias})

        # Set up new LoRA weights
        self.lora_right = nn.Parameter(torch.zeros(inp, lora_dim))
        nn.init.kaiming_uniform_(self.lora_right, a=math.sqrt(5))
        self.lora_left = nn.Parameter(torch.zeros(lora_dim, out))

    def forward(self, input):
        frozen_output = self.linear(input)
        LoRA_output = input @ self.lora_right @ self.lora_left
        return frozen_output + LoRA_output
```

LoRA Method: Implementation - Adjusting the Model

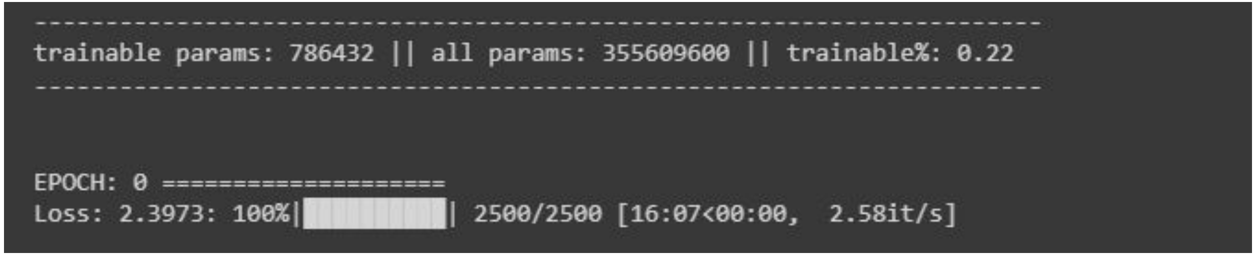


```
lora_dim = 8

# Gather target modules
targets = [n for n, _ in lora_model.named_modules() if "attn.c_attn" in n]

# replace each module with LoRA
for name in targets:
    name_struct = name.split(".")

    module_list = [lora_model]
    for struct in name_struct:
        module_list.append(getattr(module_list[-1], struct))

    # build LoRA layer
    lora = LoRA_Linear(
        weight = torch.transpose(module_list[-1].weight, 0, 1), # old weight
        bias = module_list[-1].bias, # old bias
        lora_dim = lora_dim # lora dimensionality
        )

    # set child of parent to new LoRA layer
    module_list[-2].__setattr__(name_struct[-1], lora)

# Freeze all non-LoRA params
for n, p in lora_model.named_parameters():
    p.requires_grad = "lora_right" in n or "lora_left" in n
```

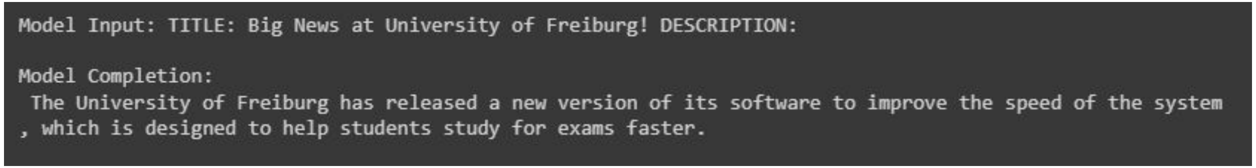
LoRA Method: Implementation - Training



```
lora_model = lora_model.to(device)

print_trainable_parameters(lora_model)
fine_tune(lora_model, epochs=1)
```



```
-----------------------------------------------------------------------
trainable params: 786432 || all params: 355609600 || trainable%: 0.22
-----------------------------------------------------------------------


EPOCH: 0 ====================
Loss: 2.3973: 100%|██████████| 2500/2500 [16:07<00:00,  2.58it/s]
```



```
Model Input: TITLE: Big News at University of Freiburg! DESCRIPTION:

Model Completion:
 The University of Freiburg has released a new version of its software to improve the speed of the system
, which is designed to help students study for exams faster.
```

Full code: https://github.com/Lukas-Liemen/Fine-tuning-gpt2-medium-with-LoRA-from-scratch

Conclusion

- LoRA is “an efficient adaptation strategy that neither introduces inference latency nor reduces input sequence length while retaining high model quality”
- Allows switching between tasks easily
- Applicable to any neural network with dense layers

- Possible future work:
  - Combination with other methods
  - Better understanding of mechanism behind fine-tuning
  - Which weights matrices to apply LoRA to (without depending on heuristics)
  - Frozen weights might be rank-deficient as well

Any questions left?