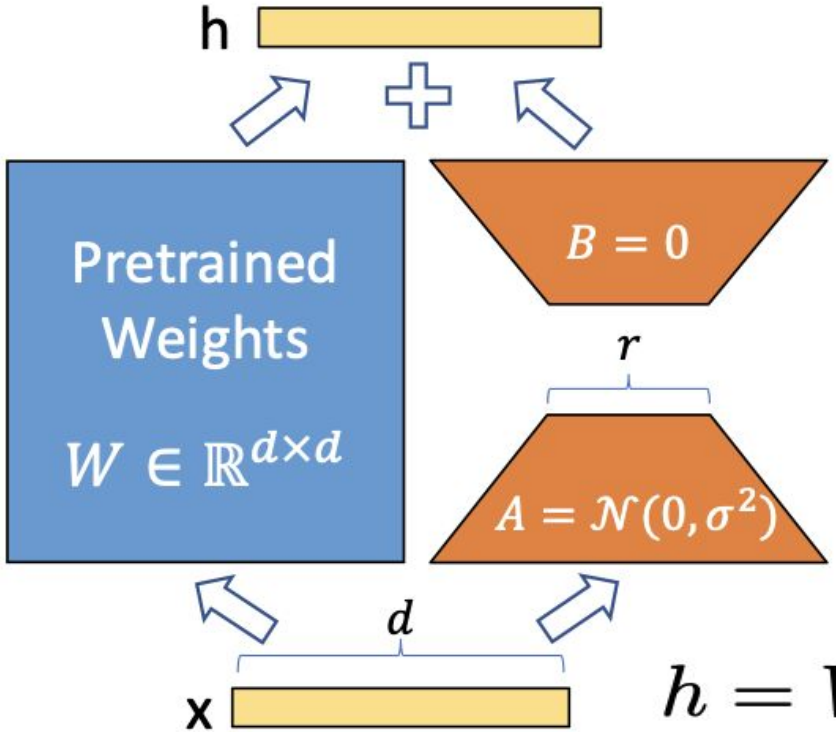
Fine-tuning project

Semenov A.

Outline

1. LoRA a brief overview
2. Llama-LoRA-Orca
3. Llama-LoRA-CustomDataset
4. Llama-LoRA-MixedDatasets
5. Llama-LoRA-TranslationDatset
6. Sum Up

LoRA



$$W_0 \in \mathbb{R}^{d \times k}$$

$$W_0 + \Delta W = W_0 + BA$$

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

$$r \ll \min(d, k)$$

$$h = W_0 x + \Delta W x = W_0 x + BAx$$

Llama-LoRA-Orca

https://github.com/Andron00e/Fine-Tuning-project/blob/main/LlamaLoRAOrca_v2.ipynb

```
LlamaForCausalLM(
  (model): LlamaModel(
    (embed_tokens): Embedding(32000, 3200, padding_idx=0)
    (layers): ModuleList(
      (0-25): 26 x LlamaDecoderLayer(
        (self_attn): LlamaAttention(
          (q_proj): Linear8bitLt(in_features=3200, out_features=3200, bias=False)
          (k_proj): Linear8bitLt(in_features=3200, out_features=3200, bias=False)
          (v_proj): Linear8bitLt(in_features=3200, out_features=3200, bias=False)
          (o_proj): Linear8bitLt(in_features=3200, out_features=3200, bias=False)
          (rotary_emb): LlamaRotaryEmbedding()
        )
        (mlp): LlamaMLP(
          (gate_proj): Linear8bitLt(in_features=3200, out_features=8640, bias=False)
          (down_proj): Linear8bitLt(in_features=8640, out_features=3200, bias=False)
          (up_proj): Linear8bitLt(in_features=3200, out_features=8640, bias=False)
          (act_fn): SiLUActivation()
        )
        (input_layernorm): LlamaRMSNorm()
        (post_attention_layernorm): LlamaRMSNorm()
      )
    )
    (norm): LlamaRMSNorm()
  )
  (lm_head): Linear(in_features=3200, out_features=32000, bias=False)
)
```

openlm-research/open_llama_3b

```
from peft import LoraConfig, get_peft_model

config = LoraConfig(
    r=8,
    lora_alpha=16,
    target_modules=[
    "q_proj",
    "v_proj",
],
    lora_dropout=0.05,
    bias="none",
    task_type="CAUSAL_LM"
)
```

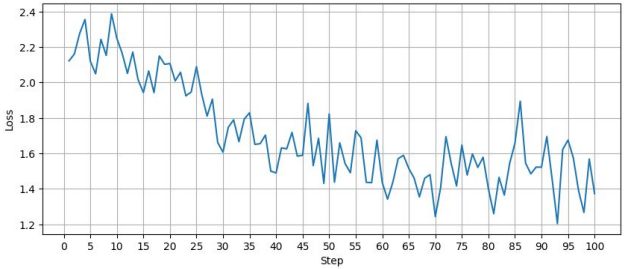
Llama-LoRA-Orca

Open-Orca/OpenOrca

```
Dataset({
    features: ['id', 'system_prompt', 'question', 'response'],
    num_rows: 42340
})
```

```
import transformers

trainer = transformers.Trainer(
    model=model,
    train_dataset=mapped_train_dataset,
    args=transformers.TrainingArguments(
        per_device_train_batch_size=1,
        gradient_accumulation_steps=16,
        warmup_steps=100,
        max_steps=100,
        learning_rate=1e-3,
        fp16=True,
        logging_steps=1,
        output_dir='outputs',
    ),
    data_collator=transformers.DataCollatorForLanguageModeling(tokenizer, mlm=False)
```



https://huggingface.co/Andron00e/YetAnother_Open-Llama-3B-LoRA-OpenOrca

Llama-LoRA-Orca

Evaluation

Llama-LoRA-Orca

Task	Version	Metric	Value		Stderr
hellaswag	0	acc	0.4899	±	0.0050
		acc_norm	0.6506	±	0.0048

CONTEXT

Cheese is the best food.

QUESTION

What is the best food?

RESPONSE

The best food is definitely cheese. It is a delicious and versatile food that can be enjoyed in many different ways. It is also a very healthy food, as it is low in calories and high in protein. Cheese is also a great source of calcium, which is essential for bone health.

CONTEXT

Teach me how to fine-tune open llama with LoRA.

QUESTION

How to fine-tune open llama with LoRA?

RESPONSE

To fine-tune open llama with LoRA, you will need to use the LoRA library to connect to the LoRa network. You can find the LoRA library in the OpenLLAMA repository. Once you have the library installed, you can use the LoRA library to connect to the LoRa network.

To connect to the LoRa network, you will need to use the LoRA library to create a LoRa device. The LoRA library provides a simple interface for creating a LoRa device. You will need to provide the device's MAC address, which is a unique identifier for the device. Once you have the device's MAC address, you can use the LoRA library to connect to the LoRa network.

Once you have connected to the LoRa network, you can use the LoRA library to fine-tune open llama. The LoRA library provides a simple interface for fine-tuning open llama. You will

Llama-LoRA-CustomDataset

Llama-LoRA-CustomDataset

```
Dataset({
    features: ['question', 'answer'],
    num_rows: 16657
})
```

QUESTION

Раскажи как надо правильно отжиматься?

ANSWER

Надо написать все в одну строку

Llama-LoRA-CustomDataset

```
1  mapped_train_dataset
```

```
Dataset({
    features: ['msg_id', 'sender', 'sender_id', 'reply_to_msg_id', 'date', 'msg_type', 'msg_content', 'input_ids', 'attention_mask'],
    num_rows: 149874
})
```

Task	Version	Metric	Value		Stderr
hellaswag	0	acc	0.4818	±	0.0050
		acc_norm	0.6377	±	0.0048

Llama-LoRA-CustomDataset

Hosted inference API

Text Generation — Example 5

Сколько у шимпанзе?

На

Compute ⌘+Enter 3.5

Hosted inference API

Text Generation — Example 5

Меня зовут Андрей и я вас жду

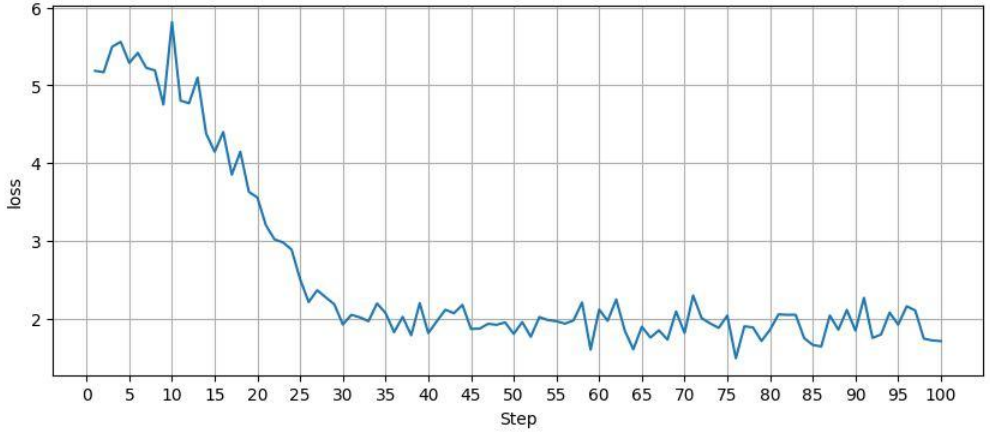
Compute ⌘+Enter 2.9

Hosted inference API

Text Generation — Example 8

Однажды, когда Джозеф

Compute ⌘+Enter 4.2



Text Generation — Example 5

My name is Lewis and I like to play football.
I am a 14 year old

Compute ⌘+Enter 4.7

Hosted inference API

Text Generation — Example 8

Once upon a time, the only way to get a job was to go to college.
But

Compute ⌘+Enter 5.4

Llama-LoRA-CustomDataset

https://github.com/Andron00e/Fine-Tuning-project/blob/main/Inference_LoRA.ipynb

Original style

CONTEXT

Your name is Sasha and you live in Moscow. What is your name?

RESPONSE

Я Дмитрий

CONTEXT

Сыр это самая вкусная еда в мире. Какая еда в мире самая вкусная?

RESPONSE

Это Большая Баллада

Q&A style

CONTEXT

Тебя зовут Саша и ты живешь в Москве

QUESTION

Как тебя зовут?

RESPONSE

Тебя зовут Саша и ты живешь в Москве

CONTEXT

Сыр это самая вкусная еда в мире.

QUESTION

Какая еда в мире самая вкусная?

RESPONSE

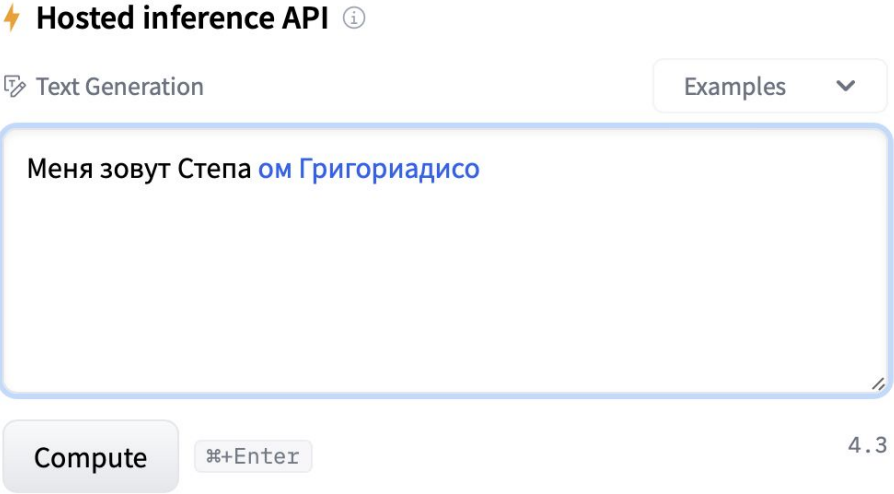
Это сыр

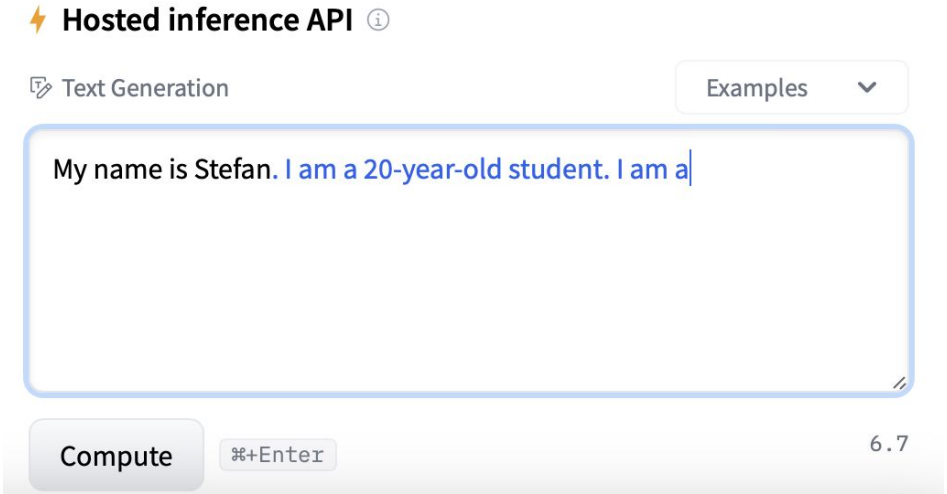
Llama-LoRA-MixedDatasets

Task	Version	Metric	Value		Stderr
hellaswag	0	acc	0.4618	±	0.0050
		acc_norm	0.6019	±	0.0049

Llama-LoRA-MixedDatasets

https://huggingface.co/Andron00e/YetAnother_Open-Llama-3B-LoRA





Llama-LoRA-TranslationDataset

https://huggingface.co/facebook/wmt19-ru-en

```
Dataset({
    features: ['input', 'output'],
    num_rows: 124974
})
```

Task	Version	Metric	Value		Stderr
hellaswag	0	acc	0.4817	±	0.0050
		acc_norm	0.6362	±	0.0048

Llama-LoRA-TranslationDataset

QUESTION

What is the most delicious dish you could ever taste?

ANSWER

Что лучше вкусное блюдо, которое можно вкусить? Например, Большой Баланс, Большой Баланс, Большой Баланс, Большой Баланс, Большой Баланс, Большой Баланс, Большой Баланс, Большой Баланс, Большой Баланс, Большой Баланс, Бо

QUESTION

Write a sequel/prequel about the 'X' movie

ANSWER

Записать последующее/предыдущее о киноленте "X" Джонатана Байрона.

QUESTION

Come up with a 14-day itinerary for a trip to Germany. The first suggested attraction should be “Take a tour of the Reichstag Building in Berlin.

ANSWER

Должно быть создано 14-дневное планирование поездки в Германию. Первый предложенный объект должен быть «Должны посетить Рейхстаг Берлин.

Sum up

1) The best q&a is fine-tuned on OpenOrca Llama 3B

2) Fine-tuned on the corpus of tg messages model not bad in q&a style but weak in text generation regime

3) Russian informal text brings more mess into model outputs and results in English has slightly worsened

4) Fine-tuned on wmt19 model is decent in translation while in text generation it outputs a lot of garbage

CONTEXT

В Арктике всегда холодно.

QUESTION

Где всегда холодно?

RESPONSE

На Байкальском кряже