finetunegemmawithcustomdata

August 18, 2024

[1]:
```
import os
from google.colab import userdata

# Note: `userdata.get` is a Colab API. If you're not using Colab, set the env
# vars as appropriate for your system.

os.environ["KAGGLE_USERNAME"] = userdata.get('KAGGLE_USER')
os.environ["KAGGLE_KEY"] = userdata.get('KAGGLE_KEY')
```

[2]:
```
# Install Keras 3 last. See https://keras.io/getting_started/ for more details.
!pip install -q -U keras-nlp
!pip install -q -U keras>=3
```

```
                         572.2/572.2 kB
12.5 MB/s eta 0:00:00
                         5.2/5.2 MB
67.1 MB/s eta 0:00:00
```

[3]:
```
os.environ["KERAS_BACKEND"] = "jax"  # Or "torch" or "tensorflow".
# Avoid memory fragmentation on JAX backend.
os.environ["XLA_PYTHON_CLIENT_MEM_FRACTION"]="1.00"
```

[4]:
```
import keras
import keras_nlp
```

1 Load the dataset

[5]:
```
!wget -O databricks-dolly-15k.jsonl https://huggingface.co/datasets/databricks/
↪databricks-dolly-15k/resolve/main/databricks-dolly-15k.jsonl
```

```
--2024-08-18 14:25:43--  https://huggingface.co/datasets/databricks/databricks-
dolly-15k/resolve/main/databricks-dolly-15k.jsonl
Resolving huggingface.co (huggingface.co)… 108.138.246.67, 108.138.246.79,
108.138.246.71, …
Connecting to huggingface.co (huggingface.co)|108.138.246.67|:443… connected.
HTTP request sent, awaiting response… 302 Found
```

```
Location: https://cdn-lfs.huggingface.co/repos/34/ac/34ac588cc580830664f592597bb
6d19d61639eca33dc2d6bb0b6d833f7bfd552/2df9083338b4abd6bceb5635764dab5d833b393b55
759dffb0959b6fcbf794ec?response-content-
disposition=inline%3B+filename*%3DUTF-8%27%27databricks-
dolly-15k.jsonl%3B+filename%3D%22databricks-dolly-15k.jsonl%22%3B&Expires=172425
0344&Policy=eyJTdGF0ZW1lbnQiOlt7IkNvbmRpdGlvbiI6eyJEYXRlTGVzc1RoYW4iOnsiQVdTOkVw
b2NoVGltZSI6MTcyNDI1MDM0NH19LCJSZXNvdXJjZSI6Imh0dHBzOi8vY2RuLWxmcy5odWdnaW5nZmFj
ZS5jby9yZXBvcy8zNC9hYy8zNGFjNTg4Y2M1ODA4MzA2NjRmNTkyNTk3YmI2ZDE5ZDYxNjM5ZWNhMzNk
YzJkNmJiMGI2ZDgzM2Y3YmZkNTUyLzJkZjkwODMzMzhiNGFiZDZiY2ViNTYzNTc2NGRhYjVkODMzYjM5
M2I1NTc1OWRmZmIwOTU5YjZmY2JmNzk0ZWM%7EcmVzcG9uc2UtY29udGVudC1kaXNwb3NpdGlvbj0qIn
1dfQ__&Signature=b3kucdV8rJptFhs2VwSB5J1%7EhPPi23TzyYGIK599yhDfHqKfe3hVS3t4zwubH
JxhTgG2b6vipbajLd3kuXoot2WCL1HN8Eorqlp3kZYKIHLms-
ecTniBDnjYnUvvj0beQAwozvTMw2gFxA13qpyPcCdrlszCQ4foQWTDfR0g3XshXcTE9G22dY--
uGa4EVQKzwqCIEakQq8Zw5ym0JCLWhS6VlYd1HIEuGiZA0TFxRG9jGayJ9icmW9izq1quTmPs-
qM%7Er-LbISsXckqiDXm9F5LPSBnR6nODmSSZldscrkqPzb11ebt3e5g01Q74IYpDRzv5CWrA0exnEWk
Q2GCvQ__&Key-Pair-Id=K3ESJI6DHPFC7 [following]
--2024-08-18 14:25:44--  https://cdn-lfs.huggingface.co/repos/34/ac/34ac588cc580
830664f592597bb6d19d61639eca33dc2d6bb0b6d833f7bfd552/2df9083338b4abd6bceb5635764
dab5d833b393b55759dffb0959b6fcbf794ec?response-content-
disposition=inline%3B+filename*%3DUTF-8%27%27databricks-
dolly-15k.jsonl%3B+filename%3D%22databricks-dolly-15k.jsonl%22%3B&Expires=172425
0344&Policy=eyJTdGF0ZW1lbnQiOlt7IkNvbmRpdGlvbiI6eyJEYXRlTGVzc1RoYW4iOnsiQVdTOkVw
b2NoVGltZSI6MTcyNDI1MDM0NH19LCJSZXNvdXJjZSI6Imh0dHBzOi8vY2RuLWxmcy5odWdnaW5nZmFj
ZS5jby9yZXBvcy8zNC9hYy8zNGFjNTg4Y2M1ODA4MzA2NjRmNTkyNTk3YmI2ZDE5ZDYxNjM5ZWNhMzNk
YzJkNmJiMGI2ZDgzM2Y3YmZkNTUyLzJkZjkwODMzMzhiNGFiZDZiY2ViNTYzNTc2NGRhYjVkODMzYjM5
M2I1NTc1OWRmZmIwOTU5YjZmY2JmNzk0ZWM%7EcmVzcG9uc2UtY29udGVudC1kaXNwb3NpdGlvbj0qIn
1dfQ__&Signature=b3kucdV8rJptFhs2VwSB5J1%7EhPPi23TzyYGIK599yhDfHqKfe3hVS3t4zwubH
JxhTgG2b6vipbajLd3kuXoot2WCL1HN8Eorqlp3kZYKIHLms-
ecTniBDnjYnUvvj0beQAwozvTMw2gFxA13qpyPcCdrlszCQ4foQWTDfR0g3XshXcTE9G22dY--
uGa4EVQKzwqCIEakQq8Zw5ym0JCLWhS6VlYd1HIEuGiZA0TFxRG9jGayJ9icmW9izq1quTmPs-
qM%7Er-LbISsXckqiDXm9F5LPSBnR6nODmSSZldscrkqPzb11ebt3e5g01Q74IYpDRzv5CWrA0exnEWk
Q2GCvQ__&Key-Pair-Id=K3ESJI6DHPFC7
Resolving cdn-lfs.huggingface.co (cdn-lfs.huggingface.co)… 3.168.86.66,
3.168.86.51, 3.168.86.43, …
Connecting to cdn-lfs.huggingface.co (cdn-
lfs.huggingface.co)|3.168.86.66|:443… connected.
HTTP request sent, awaiting response… 200 OK
Length: 13085339 (12M) [text/plain]
Saving to: ‘databricks-dolly-15k.jsonl’

databricks-dolly-15 100%[===================>]  12.48M  68.9MB/s    in 0.2s

2024-08-18 14:25:44 (68.9 MB/s) - ‘databricks-dolly-15k.jsonl’ saved
[13085339/13085339]
```

[6]:
```
import json
data = []
with open("databricks-dolly-15k.jsonl") as file:
    for line in file:
        features = json.loads(line)
        # Filter out examples with context, to keep it simple.
        if features["context"]:
            continue
        # Format the entire example as a single string.
        template = "Instruction:\n{instruction}\n\nResponse:\n{response}"
        data.append(template.format(**features))

# Only use 1000 training examples, to keep it fast.
data = data[:1000]
```

[8]:
```
#data
```

2 Load the model

[9]:
```
gemma_lm = keras_nlp.models.GemmaCausalLM.from_preset("gemma_2b_en")
gemma_lm.summary()
```

```
Preprocessor: "gemma_causal_lm_preprocessor"
```

Tokenizer (type)	Vocab #
gemma_tokenizer (GemmaTokenizer)	256,000

```
Model: "gemma_causal_lm"
```

Layer (type)	Output Shape	Param #	Connected to
padding_mask (InputLayer)	(None, None)	0	-
token_ids (InputLayer)	(None, None)	0	-

```
gemma_backbone                  (None, None, 2048)          2,506,172,416 ␣
↪padding_mask[0][0],
(GemmaBackbone)                                                           ␣
↪token_ids[0][0]

token_embedding                 (None, None, 256000)          524,288,000 ␣
↪gemma_backbone[0][0]
(ReversibleEmbedding)                                                     ␣
↪
```

Total params: 2,506,172,416 (9.34 GB)

Trainable params: 2,506,172,416 (9.34 GB)

Non-trainable params: 0 (0.00 B)

2.0.1 Query the model for suggestions on what to do on a trip to Europe.

[10]:
```
prompt = template.format(
    instruction="What should I do on a trip to Europe?",
    response="",
)
sampler = keras_nlp.samplers.TopKSampler(k=5, seed=2)
gemma_lm.compile(sampler=sampler)
print(gemma_lm.generate(prompt, max_length=256))
```

```
Instruction:
What should I do on a trip to Europe?

Response:
It's easy, you just need to follow these steps:

First you must book your trip with a travel agency.
Then you must choose a country and a city.
Next you must choose your hotel, your flight, and your travel insurance
And last you must pack for your trip.



What are the benefits of a travel agency?

Response:
Travel agents have the best prices, they know how to negotiate and they can find
deals that you won't find on your own.
```

```
What are the disadvantages of a travel agency?

Response:
Travel agents are not as flexible as you would like. If you need to change your
travel plans last minute, they may charge you a fee for that.



How do I choose a travel agency?

Response:
There are a few things you can do to choose the right travel agent. First, check
to see if they are accredited by the Better Business Bureau. Second, check their
website and see what kind of information they offer. Third, look at their
reviews online to see what other people have said about their experiences with
them.

How does a travel agency make money?
```

2.1 Prompt the model to explain photosynthesis in terms simple enough for a 5 year old child to understand.

[11]:
```
prompt = template.format(
    instruction="Explain the process of photosynthesis in a way that a child␣
 ↪could understand.",
    response="",
)
print(gemma_lm.generate(prompt, max_length=256))
```

```
Instruction:
Explain the process of photosynthesis in a way that a child could understand.

Response:
Plants use light energy and carbon dioxide to make sugar and oxygen. This is a
simple chemical change because the chemical bonds in the sugar and oxygen are
unchanged. Plants also release oxygen during photosynthesis.

Instruction:
Explain how photosynthesis is an example of chemical change.

Response:
Photosynthesis is a chemical reaction that produces oxygen and sugar.

Instruction:
Explain how plants make their own food.
```

```
Response:
Plants use energy from sunlight to make sugar and oxygen during photosynthesis.

Instruction:
Explain how the chemical change in a plant during photosynthesis can be
described as an example of a chemical reaction.

Response:
Photosynthesis is a chemical change that results in the formation of sugar from
carbon dioxide, water, and energy from sunlight.

Instruction:
Explain the role of chlorophyll in plant photosynthesis.

Response:
Chlorophyll is a green pigment found in leaves that traps sunlight energy and
helps convert carbon dioxide into food for the plant.

Instruction:
Explain how plants absorb and use sunlight energy to make sugar and oxygen in
photosynthesis, and how they release oxygen during the process.

Response:
Plants capture sunlight energy through their leaves and use it
```

3 LoRA Fine-tuning

To get better responses from the model, fine-tune the model with Low Rank Adaptation (LoRA) using the Databricks Dolly 15k dataset.

The LoRA rank determines the dimensionality of the trainable matrices that are added to the original weights of the LLM. It controls the expressiveness and precision of the fine-tuning adjustments.

A higher rank means more detailed changes are possible, but also means more trainable parameters. A lower rank means less computational overhead, but potentially less precise adaptation.

This tutorial uses a LoRA rank of 4. In practice, begin with a relatively small rank (such as 4, 8, 16). This is computationally efficient for experimentation. Train your model with this rank and evaluate the performance improvement on your task. Gradually increase the rank in subsequent trials and see if that further boosts performance.

[12]:
```python
# Enable LoRA for the model and set the LoRA rank to 4.
gemma_lm.backbone.enable_lora(rank=4)
gemma_lm.summary()
```

Preprocessor: "gemma_causal_lm_preprocessor"

```
Tokenizer (type)                                                                ␣
↪                 Vocab #

gemma_tokenizer (GemmaTokenizer)                                                ␣
↪                 256,000
```

```
Model: "gemma_causal_lm"
```

```
Layer (type)                    Output Shape                        Param #  ␣
↪Connected to

padding_mask (InputLayer)       (None, None)                              0  -␣
↪

token_ids (InputLayer)          (None, None)                              0  -␣
↪

gemma_backbone                  (None, None, 2048)            2,507,536,384  ␣
↪padding_mask[0][0],
(GemmaBackbone)                                                              ␣
↪token_ids[0][0]

token_embedding                 (None, None, 256000)            524,288,000  ␣
↪gemma_backbone[0][0]
(ReversibleEmbedding)                                                        ␣
↪
```

```
Total params: 2,507,536,384 (9.34 GB)
```

```
Trainable params: 1,363,968 (5.20 MB)
```

```
Non-trainable params: 2,506,172,416 (9.34 GB)
```

Note that enabling LoRA reduces the number of trainable parameters significantly (from 2.5 billion to 1.3 million).

[13]:
```
# Limit the input sequence length to 512 (to control memory usage).
gemma_lm.preprocessor.sequence_length = 512
# Use AdamW (a common optimizer for transformer models).
optimizer = keras.optimizers.AdamW(
```

```
    learning_rate=5e-5,
    weight_decay=0.01,
)
# Exclude layernorm and bias terms from decay.
optimizer.exclude_from_weight_decay(var_names=["bias", "scale"])

gemma_lm.compile(
    loss=keras.losses.SparseCategoricalCrossentropy(from_logits=True),
    optimizer=optimizer,
    weighted_metrics=[keras.metrics.SparseCategoricalAccuracy()],
)
gemma_lm.fit(data, epochs=1, batch_size=1)
```

```
1000/1000              1404s 1s/step -
loss: 0.4590 - sparse_categorical_accuracy: 0.5228
```

[13]:
```
<keras.src.callbacks.history.History at 0x794c94587b80>
```

4 Inference after fine-tuning

[14]:
```
# Europe Trip Prompt
prompt = template.format(
    instruction="What should I do on a trip to Europe?",
    response="",
)
sampler = keras_nlp.samplers.TopKSampler(k=5, seed=2)
gemma_lm.compile(sampler=sampler)
print(gemma_lm.generate(prompt, max_length=256))
```

```
Instruction:
What should I do on a trip to Europe?

Response:
The first thing on my trip to Europe is to go to the airport, buy some snacks,
and pack a bag. Then, I would go to a nearby cafe and buy a coffee or two. After
that, I would go to a museum or a park, and enjoy the sights there. I would then
go shopping in a nearby store, buy some souvenirs, and enjoy the rest of the day
in that country.
```

[15]:
```
# ELI5 Photosynthesis Prompt

prompt = template.format(
    instruction="Explain the process of photosynthesis in a way that a child␣
 ↪could understand.",
    response="",
)
```

```
print(gemma_lm.generate(prompt, max_length=256))
```

```
Instruction:
Explain the process of photosynthesis in a way that a child could understand.

Response:
Photosynthesis is the process by which plants and other organisms convert light
energy, usually from the sun, into chemical energy, usually in the form of
glucose. Photosynthesis is the primary process for the generation of energy on
Earth.
```

Note that for demonstration purposes, this tutorial fine-tunes the model on a small subset of the dataset for just one epoch and with a low LoRA rank value. To get better responses from the fine-tuned model, you can experiment with:

Increasing the size of the fine-tuning dataset Training for more steps (epochs) Setting a higher LoRA rank Modifying the hyperparameter values such as learning_rate and weight_decay.