

BLOOM-176B, как генерировать ей текст?

Библиотека **petals**, получить легкий доступ к генерации текста с помощью модели BLOOM-176B. Вот пример кода:

```
inputs = tokenizer('A cat in French is "', return_tensors="pt")["input_ids"].cuda()
outputs = model.generate(inputs, max_new_tokens=3)
print(tokenizer.decode(outputs[0]))
```

```
A cat in French is "chat",
time: 11.4 s (started: 2023-02-16 17:21:50 +00:00)
```

Эксперименты

Их я провел в пайтон ноутбуках. Сделал несколько коротких генераций и посмотрел, как модель может генерировать ответы на задачи из [GSM8K](#).

Важное - Генерация текста занимала много времени и уже в этот момент я думал о том, что не удастся посчитать ассигсу так как это займет слишком много времени. В ноутбуке вы можете увидеть время выполнения ячейки.

Ноутбук -> `Cot_sandbox.ipynb`

Датасет [GSM8K](#), разберитесь в том, как он устроен?

Датасет состоит из 8500 примеров задач с решениями из начальной школы. Соответственно есть два ключа - "question" и "answer" для каждого примера. Для задач написан не просто ответ, а дано рассуждение о том, как прийти к этому ответу - **chain of thought**. Это важно, потому что именно такие примеры рассуждений могут помочь модели генерировать правильные ответы на задачу.

О структуре данных:

Данные представляют собой два json файла - train, test в формате jsonl, который можно прочитать с помощью:

```
data = pd.read_json(url, lines = True)
```

Некоторые вычисления занесены в << ... >>. Финальный ответ на задачу находится в решении (ключ «answer») и находится после ####.

Предполагается, что данные, если нужно, можно упростить, убрав все подстроки вида << ... >>.

Пример:

```
{"question": "Janet\u2019s ducks lay 16 eggs per day. She  
eats three for breakfast every morning and bakes muffins for  
her friends every day with four. She sells the remainder at  
the farmers' market daily for $2 per fresh duck egg. How much  
in dollars does she make every day at the farmers' market?",  
  
"answer": "Janet sells 16 - 3 - 4 = <<16-3-4=9>>9 duck eggs a  
day.\nShe makes 9 * 2 = $<<9*2=18>>18 every day at the  
farmer\u2019s market.\n#### 18"}
```

В своих ноутбуках я буду пробовать убирать или оставлять << ... >>

Эксперименты со сравнением CoT и ансамблированного CoT на GSM8K с BLOOM-176B

На данный момент библиотека petals не позволяет мне генерировать текст и поэтому мое исследование остановилось, что было бы если бы это не произошло:

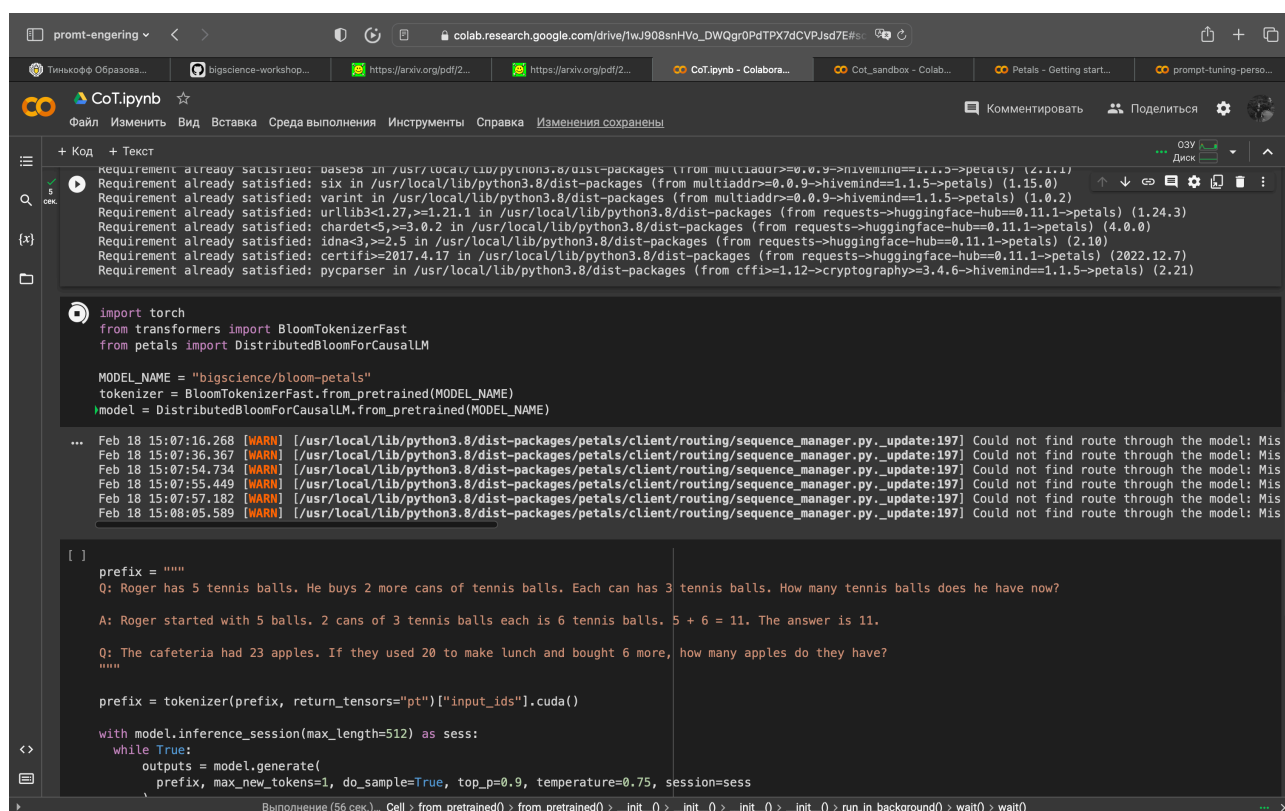
Создаем ансамблирование

Создам функцию для генерации нескольких вариантов ответа модели по входным данным с помощью изменения параметров генерации или даже изменением промптинга. Потом просто выбираем ответ, который получался максимальное количество раз.

Проводим эксперименты

Фиксируем значения и проводим сравнения с другими результатами

Проблемы:



The screenshot shows a Google Colab notebook titled 'CoT.ipynb'. The interface includes a top bar with navigation icons and tabs for 'Файл', 'Изменить', 'Вид', 'Вставка', 'Среда выполнения', 'Инструменты', 'Справка', and 'Изменения сохранены'. The main area is divided into a code editor and an output area. The code editor contains the following Python code:

```
import torch
from transformers import BloomTokenizerFast
from petals import DistributedBloomForCausalLM

MODEL_NAME = "bigscience/bloom-petals"
tokenizer = BloomTokenizerFast.from_pretrained(MODEL_NAME)
model = DistributedBloomForCausalLM.from_pretrained(MODEL_NAME)
```

The output area shows a series of warnings from the petals library, indicating that the model could not find a route through the model. The warnings are as follows:

```
Feb 18 15:07:16.268 [WARN] [/usr/local/lib/python3.8/dist-packages/petals/client/routing/sequence_manager.py._update:197] Could not find route through the model: Mis
Feb 18 15:07:36.367 [WARN] [/usr/local/lib/python3.8/dist-packages/petals/client/routing/sequence_manager.py._update:197] Could not find route through the model: Mis
Feb 18 15:07:54.734 [WARN] [/usr/local/lib/python3.8/dist-packages/petals/client/routing/sequence_manager.py._update:197] Could not find route through the model: Mis
Feb 18 15:07:55.449 [WARN] [/usr/local/lib/python3.8/dist-packages/petals/client/routing/sequence_manager.py._update:197] Could not find route through the model: Mis
Feb 18 15:07:57.182 [WARN] [/usr/local/lib/python3.8/dist-packages/petals/client/routing/sequence_manager.py._update:197] Could not find route through the model: Mis
Feb 18 15:08:05.589 [WARN] [/usr/local/lib/python3.8/dist-packages/petals/client/routing/sequence_manager.py._update:197] Could not find route through the model: Mis
```

Below the warnings, the code continues with a prefix and a question:

```
[ ]
prefix = ""
Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. 5 + 6 = 11. The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

prefix = tokenizer(prefix, return_tensors="pt")["input_ids"].cuda()

with model.inference_session(max_length=512) as sess:
    while True:
        outputs = model.generate(
            prefix, max_new_tokens=1, do_sample=True, top_p=0.9, temperature=0.75, session=sess
```

Feb 19 14:04:35.459 [WARN] [/usr/local/lib/python3.8/dist-packages/petals/client/routing/sequence_manager.py._update:197] Could not find route through the model: MissingBlocksError("No servers holding blocks [28, 29, 30] are online.\nYou can check the public swarm's state at <http://health.petals.ml>\n\nIf there are not enough servers, please consider connecting your own GPU:\nhttps://github.com/bigscience-workshop/petals#connect-your-gpu-and-increase-petals-capacity") (retry in 1 sec)

Feb 19 14:04:42.957 [WARN] [/usr/local/lib/python3.8/dist-packages/petals/client/routing/sequence_manager.py._update:197] Could not find route through the model: MissingBlocksError("No servers holding blocks [28, 29, 30] are online.\nYou can check the public swarm's state at <http://health.petals.ml>\n\nIf there are not enough servers, please consider connecting your own GPU:\nhttps://github.com/bigscience-workshop/petals#connect-your-gpu-and-increase-petals-capacity") (retry in 2 sec)

Feb 19 14:04:45.920 [WARN] [/usr/local/lib/python3.8/dist-packages/petals/client/routing/sequence_manager.py._update:197] Could not find route through the model: MissingBlocksError("No servers holding blocks [28, 29, 30] are online.\nYou can check the public swarm's state at <http://health.petals.ml>\n\nIf there are not enough servers, please consider connecting your own GPU:\nhttps://github.com/bigscience-workshop/petals#connect-your-gpu-and-increase-petals-capacity") (retry in 4 sec)

Feb 19 14:04:56.574 [WARN] [/usr/local/lib/python3.8/dist-packages/petals/client/routing/sequence_manager.py._update:197] Could not find route through the model:

```
MissingBlocksError("No servers holding blocks [28, 29, 30] are online.\nYou can check the  
public swarm's state at http://health.petals.ml\n\nIf there are not enough servers,  
please consider connecting your own GPU:\nhttps://github.com/bigscience-workshop/  
petals#connect-your-gpu-and-increase-petals-capacity") (retry in 8 sec)  
Feb 19 14:05:05.506 [WARN] [/usr/local/lib/python3.8/dist-packages/petals/client/routing/  
sequence_manager.py. _update:197] Could not find route through the model:  
MissingBlocksError("No servers holding blocks [28, 29, 30] are online.\nYou can check the  
public swarm's state at http://health.petals.ml\n\nIf there are not enough servers,  
please consider connecting your own GPU:\nhttps://github.com/bigscience-workshop/  
petals#connect-your-gpu-and-increase-petals-capacity") (retry in 16 sec)
```