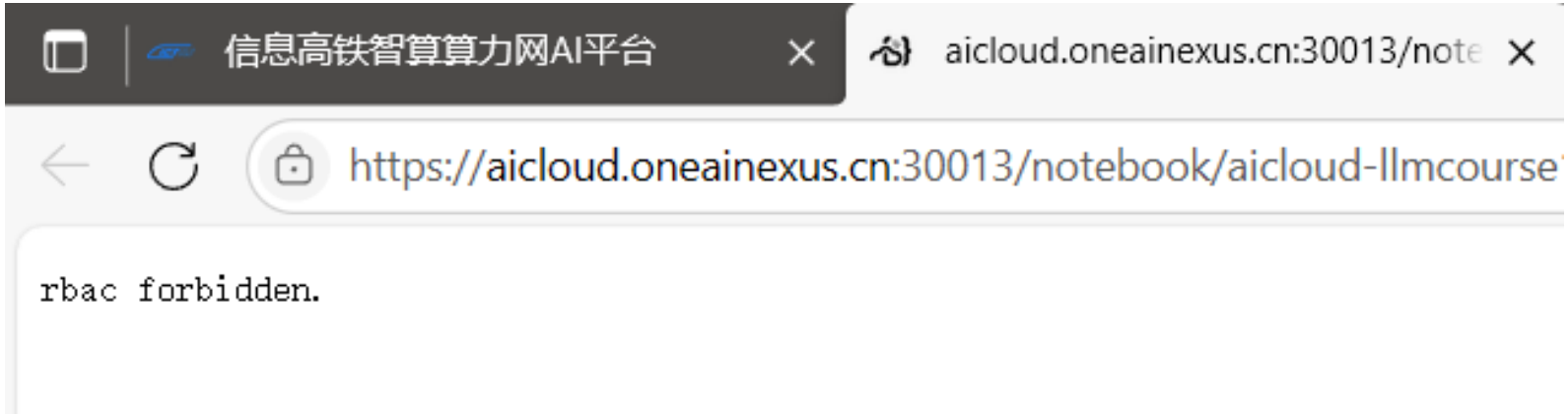
通用大模型原理及训练实践
实验课③：答疑

中国科学院计算技术研究所
Institute of Computing Technology,Chinese Academy of Sciences

连接服务器时"rabc forbidden"

- 关闭标签页
- 重新连接，若还不行则刷新信息高铁算力平台主页，若还不行则重新登录



指令数据格式

- 无论是哪种指令数据（通用问答/特定任务/自我认知），都是<instruction, input, output>三元组格式的数据
  - instruction是用户的指令
  - input是用户输入，该字段可以为空
  - output是系统输出
- 训练集为单个json文件，包含若干条目，每个条目为上述三元组

不同数据合并

- 三部分数据合并到一个json文件中
- 如果某部分数据过少，可以采取复制多份的策略，防止其在训练过程中被忽视

```python
import json
import argparse

parser = argparse.ArgumentParser()
parser.add_argument("--input-files")
parser.add_argument("--output-path")
args = parser.parse_args()

file_list = args.input_files.split(",")
merge_data = []
for filename in file_list:
    with open (filename, "r") as f:
        data = json.load(f)
        merge_data.extend(data)
with open(args.output_path, "w", encoding="utf-8") as f:
    json.dump(merge_data, f, indent=4, ensure_ascii=False)
```

A100机器8bit训练

- 报错：ValueError: `.to` is not supported for `4-bit` or `8-bit` bitsandbytes models. Please use the model as it is, since the model has already been set to the correct devices and casted to the correct `dtype`.
- 解决办法：先更新transformers: pip install -U transformers。更新后需要修改trainer = transformers.Trainer这部分的参数，其中args=transformers.TrainingArguments里面的evaluation_strategy="steps"删去，save_strategy改为"no"，如图：

```
trainer = transformers.Trainer(
    model=model,
    train_dataset=train_data,
    eval_dataset=val_data,
    args=transformers.TrainingArguments(
        per_device_train_batch_size=micro_batch_size,
        gradient_accumulation_steps=gradient_accumulation_steps,
        warmup_steps=100,
        num_train_epochs=num_epochs,
        learning_rate=learning_rate,
        fp16=True,
        logging_steps=10,
        optim="adamw_torch",
        # evaluation_strategy="steps" if val_set_size > 0 else "no",
        save_strategy="no",
        eval_steps=200 if val_set_size > 0 else None,
        save_steps=200,
        output_dir=output_dir,
        save_total_limit=3,
        load_best_model_at_end=True if val_set_size > 0 else False,
        ddp_find_unused_parameters=False if ddp else None,
        group_by_length=group_by_length,
        report_to="wandb" if use_wandb else None,
        run_name=wandb_run_name if use_wandb else None,
    ),
```

cutoff_len与micro_batch_size

- cutoff_len：截断长度
- micro_batch_size：单次前向计算的batch size

- 如果指令数据较长，总是被截断，可以调大cutoff_len
- 如果出现CUDA: out of memory，可以调小micro_batch_size

- A100 40G机器，cutoff_len=512时micro_batch_size能开到

模型导出错误

```
Traceback (most recent call last):
  File "export_hf_checkpoint.py", line 47, in <module>
    assert not torch.allclose(first_weight_old, first_weight)
RuntimeError: Half did not match Float
```

- 手动将模型参数转为 fp16

```
lora_model.train(False).half()

# did we do anything?
assert not torch.allclose(first_weight_old,
first_weight)
```

chat.py启动时报错

- 原因1：模型路径需包含bayling，否则会报错，请确认路径无误
- 原因2：导出的模型默认不包含以下3个文件：
  - tokenizer.model
  - tokenizer_config.json
  - tokenizer.json
  - 将bayling-2-7b目录下的以上3个文件复制过去即可

谢谢！