通过语言对齐将大语言模型英语能力外推到非英语语言

1. 配置环境

1.1 按照README.md安装库存在的问题

> conda env create -f environment.yml

1. 会长期卡在 Installing pip dependencies：

   尝试对environment.yml文件进行以下修改，添加镜像源即可：
   将channels改为（注意要把default去掉）：

```
channels:
  - conda-forge
  - https://mirrors.tuna.tsinghua.edu.cn/anaconda/pkgs/main
  - https://mirrors.tuna.tsinghua.edu.cn/anaconda/pkgs/free
  - https://mirrors.tuna.tsinghua.edu.cn/anaconda/pkgs/r
  - https://mirrors.tuna.tsinghua.edu.cn/anaconda/pkgs/pro
  - https://mirrors.tuna.tsinghua.edu.cn/anaconda/pkgs/msys2
```

   并在pip的依赖包里添加上镜像源（加上最后一行）。

```
  - pip:
    - addict==2.4.0
    - anyio==3.3.0
        .......
    - websocket-client==1.1.0
    - widgetsnbextension==3.5.1
    - sapien==1.1.1
    - -i https://pypi.tuna.tsinghua.edu.cn/simple
```

2. 找不到包满足bleurt==0.0.2

   操作：删除bleurt并自行安装（去除版本限制也会报错，找不到对应的包）

   方法：参照google-research/bleurt: BLEURT is a metric for Natural Language Generation based on transfer learning. (github.com)手动安装

```
pip install --upgrade pip  # ensures that pip is current
git clone https://github.com/google-research/bleurt.git
cd bleurt
pip install .
```

3. 安装tensorrt-libs==8.6.1报错

   原因：删除tensorrt-libs==8.6.1并自行安装（去除版本限制也会报错，子进程报错）

   方法：再次使用 pip install tensorrt-libs==8.6.1会发现已经安装了

4. 存在包版本错误冲突问题：

> The conflict is caused by:
>
> The user requested typing-extensions==4.7.1
>
> altair 5.0.1 depends on typing-extensions>=4.0.1; python_version < "3.11"
>
> fastapi 0.101.0 depends on typing-extensions>=4.5.0
>
> gradio 3.39.0 depends on typing-extensions~=4.0
>
> gradio-client 0.3.0 depends on typing-extensions~=4.0
>
> huggingface-hub 0.16.4 depends on typing-extensions>=3.7.4.3
>
> lightning-utilities 0.8.0 depends on typing-extensions
>
> pydantic 2.1.1 depends on typing-extensions>=4.6.1
>
> pydantic-core 2.4.0 depends on typing-extensions!=4.7.0 and >=4.6.0
>
> pyre-extensions 0.0.29 depends on typing-extensions
>
> pytorch-lightning 1.9.5 depends on typing-extensions>=4.0.0
>
> tensorflow 2.13.0 depends on typing-extensions<4.6.0 and >=3.6.6
>
> To fix this you could try to:
>
> 1. loosen the range of package versions you've specified
> 2. remove package versions to allow pip attempt to solve the dependency conflict
>
> Pip subprocess error:
>
> ERROR: Cannot install -r /home/djh/code/xllm/condaenv.dxptyxf0.requirements.txt (line 121), -r /home/djh/code/xllm/condaenv.dxptyxf0.requirements.txt (line 28), -r /home/djh/code/xllm/condaenv.dxptyxf0.requirements.txt (line 41), -r /home/djh/code/xllm/condaenv.dxptyxf0.requirements.txt (line 42), -r

> /home/djh/code/xllm/condaenv.dxptyxf0.requirements.txt (line 48), -r /home/djh/code/xllm/condaenv.dxptyxf0.requirements.txt (line 56), -r /home/djh/code/xllm/condaenv.dxptyxf0.requirements.txt (line 6), -r /home/djh/code/xllm/condaenv.dxptyxf0.requirements.txt (line 90), -r /home/djh/code/xllm/condaenv.dxptyxf0.requirements.txt (line 91), -r /home/djh/code/xllm/condaenv.dxptyxf0.requirements.txt (line 95), -r /home/djh/code/xllm/condaenv.dxptyxf0.requirements.txt (line 97) and typing-extensions==4.7.1 because these package versions have conflicting dependencies.
>
> ERROR: ResolutionImpossible: for help visit https://pip.pypa.io/en/latest/topics/dependency-resolution/#dealing-with-dependency-conflicts
>
> failed
>
> CondaEnvException: Pip failed

这个冲突是由于以下原因引起的：

- 用户请求了 `typing-extensions==4.7.1`
- `altair 5.0.1` 依赖于 `typing-extensions>=4.0.1; python_version <`
- `fastapi 0.101.0` 依赖于 `typing-extensions>=4.5.0`
- `gradio 3.39.0` 依赖于 `typing-extensions~=4.0`
- `gradio-client 0.3.0` 依赖于 `typing-extensions~=4.0`
- `huggingface-hub 0.16.4` 依赖于 `typing-extensions>=3.7.4.3`
- `lightning-utilities 0.8.0` 依赖于 `typing-extensions`
- `pydantic 2.1.1` 依赖于 `typing-extensions>=4.6.1`
- `pydantic-core 2.4.0` 依赖于 `typing-extensions!=4.7.0 and >=4.6.0`
- `pyre-extensions 0.0.29` 依赖于 `typing-extensions`
- `pytorch-lightning 1.9.5` 依赖于 `typing-extensions>=4.0.0`
- `tensorflow 2.13.0` 依赖于 `typing-extensions<4.6.0 and >=3.6.6`

为了解决这个问题，您可以尝试以下方法：

1. 放宽您指定的软件包版本范围。
2. 删除软件包版本，以便允许 pip 尝试解决依赖冲突。

- 首先尝试去掉 `tensorflow` 的包版本限制

conda env update -f environment.yml

然后会报类似的错误，依次取消upbabel-comet==2.0.1的限制、tensorboard==2.13.0 typing-extensions==4.7.1 keras==2.13.1 wrapt==1.15.0 google-auth-oauthlib==1.0.0 tensorboard-data-server==0.7.1 google-auth==2.23.0

报错没有尽头

另一种方式：

原因：考虑到typing-extensions(==4.7.1) 但多个其他包依赖不同的typing-extensions版本

操作：openai==0.27.7需要自行安装（具体内部原因不明）

1.2 修改environment.yml后继续安装存在的问题

1. 去掉了pip后面所有包的版本号，同时根据requirements.txt的要求保留了

```
numpy
rouge_score
fire
openai
transformers>=4.28.1
torch
sentencepiece
tokenizers>=0.13.3
wandb
```

2. 需要和本地cuda环境匹配的pytorch

```
conda install pytorch==2.0.1 torchvision==0.15.2 torchaudio==2.0.2 pytorch-cuda=11.7 -c pytorch -c nvidia
```

3. 需要重新安装utils
4. 需要重新安装openai==0.27.7
5. 需要再次重新更新包

```
conda env update -f environment.yml
```

6. 重复2-4
7. 删除bleurt并自行安装

操作：删除bleurt并自行安装（去除版本限制也会报错，找不到对应的包）

方法：参照google-research/bleurt: BLEURT is a metric for Natural Language Generation based on transfer learning. (github.com)手动安装

```
pip install --upgrade pip  # ensures that pip is current
git clone https://github.com/google-research/bleurt.git
cd bleurt
pip install .
```

8. 删除tensorrt-libs并自行安装

原因：删除tensorrt-libs==8.6.1并自行安装（去除版本限制也会报错，子进程报错）

方法：再次使用 pip install tensorrt-libs==8.6.1会发现已经安装了

运行 `bash script/train.sh llama-7b-hf alpaca_en+alpaca_zh+translation_ncwm_en-zh` 中：

> WARNING:root:Formatting inputs... 格式化输入...
>
> WARNING:root:Tokenizing inputs... This may take some time... 分词输入... 这可能需要一些时间...

报错：

> /home/djh/miniconda3/envs/xllm2/lib/python3.10/site-packages/torch/distributed/fsdp/_init_utils.py:295: UserWarning: FSDP is switching to use `NO_SHARD` instead of ShardingStrategy.FULL_SHARD since the world size is 1.
>
> warnings.warn(
>
> Traceback (most recent call last):
>
> File "/home/djh/code/xllm/train.py", line 326, in <module>
>
> train()
>
> File "/home/djh/code/xllm/train.py", line 318, in train
>
> trainer.train()
>
> File "/home/djh/miniconda3/envs/xllm2/lib/python3.10/site-packages/transformers/trainer.py", line 1664, in train
>
> return inner_training_loop(
>
> File "/home/djh/miniconda3/envs/xllm2/lib/python3.10/site-packages/transformers/trainer.py", line 1759, in _inner_training_loop
>
> model = self._wrap_model(self.model_wrapped)
>
> File "/home/djh/miniconda3/envs/xllm2/lib/python3.10/site-packages/transformers/trainer.py", line 1490, in _wrap_model
>
> self.model = model = FSDP(

File "/home/djh/miniconda3/envs/xllm2/lib/python3.10/site-packages/torch/distributed/fsdp/fully_sharded_data_parallel.py", line 408, in __init__

_init_param_handle_from_module(

File "/home/djh/miniconda3/envs/xllm2/lib/python3.10/site-packages/torch/distributed/fsdp/_init_utils.py", line 415, in _init_param_handle_from_module

_move_module_to_device(

File "/home/djh/miniconda3/envs/xllm2/lib/python3.10/site-packages/torch/distributed/fsdp/_init_utils.py", line 802, in _move_module_to_device

module = module.to(device_from_device_id)

File "/home/djh/miniconda3/envs/xllm2/lib/python3.10/site-packages/transformers/modeling_utils.py", line 1886, in to

return super().to(*args, **kwargs)

File "/home/djh/miniconda3/envs/xllm2/lib/python3.10/site-packages/torch/nn/modules/module.py", line 1145, in to

return self._apply(convert)

File "/home/djh/miniconda3/envs/xllm2/lib/python3.10/site-packages/torch/nn/modules/module.py", line 797, in _apply

module._apply(fn)

File "/home/djh/miniconda3/envs/xllm2/lib/python3.10/site-packages/torch/nn/modules/module.py", line 797, in _apply

module._apply(fn)

File "/home/djh/miniconda3/envs/xllm2/lib/python3.10/site-packages/torch/nn/modules/module.py", line 820, in _apply

param_applied = fn(param)

File "/home/djh/miniconda3/envs/xllm2/lib/python3.10/site-packages/torch/nn/modules/module.py", line 1143, in convert

return t.to(device, dtype if t.is_floating_point() or t.is_complex() else None, non_blocking)

RuntimeError: CUDA error: device kernel image is invalid

CUDA kernel errors might be asynchronously reported at some other API call, so the stacktrace below might be incorrect.

For debugging consider passing CUDA_LAUNCH_BLOCKING=1.

Compile with `TORCH_USE_CUDA_DSA` to enable device-side assertions.

ERROR:torch.distributed.elastic.multiprocessing.api:failed (exitcode: 1) local_rank: 0 (pid: 308066) of binary: /home/djh/miniconda3/envs/xllm2/bin/python

Traceback (most recent call last):

File "/home/djh/miniconda3/envs/xllm2/bin/torchrun", line 33, in <module>

sys.exit(load_entry_point('torch==2.0.1', 'console_scripts', 'torchrun')())

File "/home/djh/miniconda3/envs/xllm2/lib/python3.10/site-packages/torch/distributed/elastic/multiprocessing/errors/__init__.py", line 346, in wrapper

return f(*args, **kwargs)

File "/home/djh/miniconda3/envs/xllm2/lib/python3.10/site-packages/torch/distributed/run.py", line 794, in main

run(args)

File "/home/djh/miniconda3/envs/xllm2/lib/python3.10/site-packages/torch/distributed/run.py", line 785, in run

elastic_launch(

File "/home/djh/miniconda3/envs/xllm2/lib/python3.10/site-packages/torch/distributed/launcher/api.py", line 134, in __call__

return launch_agent(self._config, self._entrypoint, list(args))

File "/home/djh/miniconda3/envs/xllm2/lib/python3.10/site-packages/torch/distributed/launcher/api.py", line 250, in launch_agent

raise ChildFailedError(

torch.distributed.elastic.multiprocessing.errors.ChildFailedError:

==

/home/djh/code/xllm/train.py FAILED

Failures:

<NO_OTHER_FAILURES>

Root Cause (first observed failure):

[0]:

time : 2024-03-15_11:59:58

host : djh-PowerEdge-T640

> rank : 0 (local_rank: 0)
>
> exitcode : 1 (pid: 308066)
>
> error_file: <N/A>
>
> traceback : To enable traceback see: https://pytorch.org/docs/stable/elastic/errors.html
>
> ==

CUDA错误 - 设备内核映像无效:

- 这通常意味着PyTorch试图在不支持的CUDA版本上运行操作，或者CUDA设备与当前的PyTorch或CUDA版本不兼容。确保您的CUDA版本与安装的PyTorch版本兼容。
- 考虑不改变cuda版本的情况下，能否找到适应的pytorch版本
- 可以参考 硬件驱动有关问题
  - conda install python=3.10.12
    conda install pytorch=2.0.1 torchvision torchaudio pytorch-cuda=11.7 -c pytorch -c nvidia
- 目前python为3.10.11，cuda为11.7
  - pytorch历史版本参照：https://pytorch.org/get-started/previous-versions/
  - 存在报错

```
# CUDA 11.7
conda install pytorch==2.0.1 torchvision==0.15.2 torchaudio==2.0.2
pytorch-cuda=11.7 -c pytorch -c nvidia
```

  - 存在报错

```
# CUDA 11.7
conda install pytorch==2.0.0 torchvision==0.15.0 torchaudio==2.0.0
pytorch-cuda=11.7 -c pytorch -c nvidia
```

  - 存在报错

```
# CUDA 11.7
conda install pytorch==1.13.1 torchvision==0.14.1 torchaudio==0.13.1
pytorch-cuda=11.7 -c pytorch -c nvidia
```

  - 核心错误还是：cuda/torch/nvidia硬件版本过低造成的

1.3 自行手动从前向后安装

1. 考虑到硬件要求：硬件驱动有关问题，参照其中内容完成python和pytorch的安装
    a. FSDP要求pytorch必须>=2.1.0:
2. conda env update -f environment.yml（无版本号模式）
3. conda env update -f environment.yml(有版本号就一定会卡在1.1里面提过的第4点包管理冲突上)
4. Failed to initialize NVML: Driver/library version mismatch-CSDN博客

参考版本：

a. pip install openai==0.27.7
b. pip install transformers==4.29.0
c. pip install datasets==2.12.0
d. pip install openai==0.27.7
e. pip install accelerate==0.19.0
f. pip install sentencepiece==0.1.99
g. pip install -r requirements.txt
h. conda env update -f environment.yml
i. pip install
j. altair==5.0.1
k. fastapi==0.101.0
l. gradio==3.39.0
m. gradio-client==0.3.0
n. huggingface-hub==0.16.4
o. lightning-utilities==0.8.0
p. pydantic==2.1.1
q. pydantic-core==2.4.0
r. pyre-extensions==0.0.29
s. pytorch-lightning==1.9.5
t. tensorflow==2.13.0