Issues with AudioLottery code

May 5, 2024

Critical Issue: Some of the libraries they used were optimized for Linux, which is why they implemented Kubernetes. Kubernetes allows for the compartmentalization of code and databases, running them on virtual machines. However, we faced challenges replicating their setup because installing Kubernetes and Docker on our system proved to be a complex process. Therefore, we used Google Colab but we faced the following issues.

We tried to replicate the Conformer part of AudioLottery on Google colab and we encountered the following specific issues. (Link to Colab notebook)

1 Problem installing 'ctcdecode' module

We ran the following lines of code to install the module.

```
# get the code
! git clone --recursive https://github.com/parlance/ctcdecode.git
%cd ctcdecode
! pip install .
```

This gave us the following error:

1.1 Fix

'ctcdecode' needs torch 1.11.0 while we were using torch 2.3.0. Thus, the following line fixes the error:

```
! pip install torch==1.11.0
```

2 Problem with 'warp-rnnt' module

When we tried to run the main_lth.py file, we got the the following error:

```
# get the code
! git clone --recursive https://github.com/parlance/ctcdecode.git
! cd ctcdecode && pip install .
```

```
Cloning into 'ctcdecode'...
remote: Enumerating objects: 1102, done.
remote: Counting objects: 100% (39/39), done.
remote: Compressing objects: 100% (25/25), done.
remote: Total 1102 (delta 16), reused 32 (delta 14), pack-reused 1063
Receiving objects: 100% (1102/1102), 782.27 KiB | 2.90 MiB/s, done.
Resolving deltas: 100% (529/529), done.
Submodule 'third_party/ThreadPool' (https://github.com/progschj/ThreadPool.git) registered for path 'third_party/ThreadPool
Submodule 'third_party/kenlm' (https://github.com/kpu/kenlm.git) registered for path 'third_party/kenlm'
Cloning into '/content/Conformer/ctcdecode/third_party/ThreadPool'...
remote: Enumerating objects: 82, done.
remote: Counting objects: 100% (26/26), done.
remote: Compressing objects: 100% (9/9), done.
remote: Total 82 (delta 19), reused 17 (delta 17), pack-reused 56
Receiving objects: 100% (82/82), 13.34 KiB | 284.00 KiB/s, done.
Resolving deltas: 100% (36/36), done.
Cloning into '/content/Conformer/ctcdecode/third_party/kenlm'...
remote: Enumerating objects: 14165, done.
remote: Counting objects: 100% (478/478), done.
remote: Compressing objects: 100% (332/332), done.
remote: Total 14165 (delta 163), reused 409 (delta 132), pack-reused 13687
Receiving objects: 100% (14165/14165), 5.91 MiB | 10.83 MiB/s, done.
Resolving deltas: 100% (8043/8043), done.
Submodule path 'third_party/ThreadPool': checked out '9a42ec1329f259a5f4881a291db1dcb8f2ad9040'
Submodule path 'third_party/kenlm': checked out '35835f1ac4884126458ac89f9bf6dd9ccad561e0'
Processing /content/Conformer/ctcdecode
  Preparing metadata (setup.py) ... done
Building wheels for collected packages: ctcdecode
  error: subprocess-exited-with-error

  × python setup.py bdist_wheel did not run successfully.
  │ exit code: 1
  ╰─> See above for output.

  note: This error originates from a subprocess, and is likely not a problem with pip.
  Building wheel for ctcdecode (setup.py) ... error
  ERROR: Failed building wheel for ctcdecode
  Running setup.py clean for ctcdecode
Failed to build ctcdecode
ERROR: Could not build wheels for ctcdecode, which is required to install pyproject.toml-based projects
```

Figure 1: Error

```
[ ] # get the code
    ! git clone --recursive https://github.com/parlance/ctcdecode.git
```

```
Cloning into 'ctcdecode'...
remote: Enumerating objects: 1102, done.
remote: Counting objects: 100% (39/39), done.
remote: Compressing objects: 100% (25/25), done.
remote: Total 1102 (delta 16), reused 32 (delta 14), pack-reused 1063
Receiving objects: 100% (1102/1102), 782.27 KiB | 6.63 MiB/s, done.
Resolving deltas: 100% (529/529), done.
Submodule 'third_party/ThreadPool' (https://github.com/progschj/ThreadPool.git) registered for path 'third_party/ThreadPool'
Submodule 'third_party/kenlm' (https://github.com/kpu/kenlm.git) registered for path 'third_party/kenlm'
Cloning into '/content/Conformer/ctcdecode/third_party/ThreadPool'...
remote: Enumerating objects: 82, done.
remote: Counting objects: 100% (26/26), done.
remote: Compressing objects: 100% (9/9), done.
remote: Total 82 (delta 19), reused 17 (delta 17), pack-reused 56
Receiving objects: 100% (82/82), 13.34 KiB | 6.67 MiB/s, done.
Resolving deltas: 100% (36/36), done.
Cloning into '/content/Conformer/ctcdecode/third_party/kenlm'...
remote: Enumerating objects: 14165, done.
remote: Counting objects: 100% (478/478), done.
remote: Compressing objects: 100% (331/331), done.
remote: Total 14165 (delta 163), reused 409 (delta 133), pack-reused 13687
Receiving objects: 100% (14165/14165), 5.91 MiB | 16.72 MiB/s, done.
Resolving deltas: 100% (8043/8043), done.
Submodule path 'third_party/ThreadPool': checked out '9a42ec1329f259a5f4881a291db1dcb8f2ad9040'
Submodule path 'third_party/kenlm': checked out '35835f1ac4884126458ac89f9bf6dd9ccad561e0'
```

```
[ ] %cd ctcdecode
    ! pip install .
```

```
/content/Conformer/ctcdecode
Processing /content/Conformer/ctcdecode
  Preparing metadata (setup.py) ... done
Building wheels for collected packages: ctcdecode
  Building wheel for ctcdecode (setup.py) ... done
  Created wheel for ctcdecode: filename=ctcdecode-1.0.3-cp310-cp310-linux_x86_64.whl size=18914754 sha256=db6eca172a495b6f54446cf3bb7203f3ae282ba22902cce52fb9123caaebf476
  Stored in directory: /tmp/pip-ephem-wheel-cache-fbk63k15/wheels/1d/ff/65/95ad5d4f56de1249a32f1ea1fbef21e6a11476518ceb7d2212
Successfully built ctcdecode
Installing collected packages: ctcdecode
Successfully installed ctcdecode-1.0.3
```

Figure 2: Fix

```
[ ] !python main_lth.py --config_file --prepare_dataset --create_tokenizer configs/EfficientConformerCTCLargeLTH.json

    2024-05-03 06:36:38.220873: E external/local_xla/xla/stream_executor/cuda/cuda_dnn.cc:9261] Unable to register cuDNN factory: Attempting to register factory for plugin cuDNN when one has already been registered
    2024-05-03 06:36:38.220933: E external/local_xla/xla/stream_executor/cuda/cuda_fft.cc:607] Unable to register cuFFT factory: Attempting to register factory for plugin cuFFT when one has already been registered
    2024-05-03 06:36:38.375952: E external/local_xla/xla/stream_executor/cuda/cuda_blas.cc:1515] Unable to register cuBLAS factory: Attempting to register factory for plugin cuBLAS when one has already been registered
    2024-05-03 06:36:38.650029: I tensorflow/core/platform/cpu_feature_guard.cc:182] This TensorFlow binary is optimized to use available CPU instructions in performance-critical operations.
    To enable the following instructions: AVX2 FMA, in other operations, rebuild TensorFlow with the appropriate compiler flags.
    2024-05-03 06:36:40.908854: W tensorflow/compiler/tf2tensorrt/utils/py_utils.cc:38] TF-TRT Warning: Could not find TensorRT
    /content/ConFormer/models/modules.py:420: SyntaxWarning: "is not" with a literal. Did you mean "!="?
      assert not (group_size > 1 and stride > 1 is not None), "Strided grouped attention not implemented"
    Traceback (most recent call last):
      File "/content/Conformer/main_lth.py", line 5, in <module>
        from functions import *
      File "/content/Conformer/functions.py", line 5, in <module>
        from models.transducer import Transducer
      File "/content/Conformer/models/transducer.py", line 40, in <module>
        from models.lm import (
      File "/content/Conformer/models/lm.py", line 29, in <module>
        from models.losses import (
      File "/content/Conformer/models/losses.py", line 20, in <module>
        import warp_rnnt
      File "/usr/local/lib/python3.10/dist-packages/warp_rnnt/__init__.py", line 2, in <module>
        import warp_rnnt._C as core
    ImportError: /usr/local/lib/python3.10/dist-packages/warp_rnnt/_C.cpython-310-x86_64-linux-gnu.so: undefined symbol: _ZN3c104cuda14ExchangeDeviceEa
```

Figure 3: Fix