HyperDex User Guide

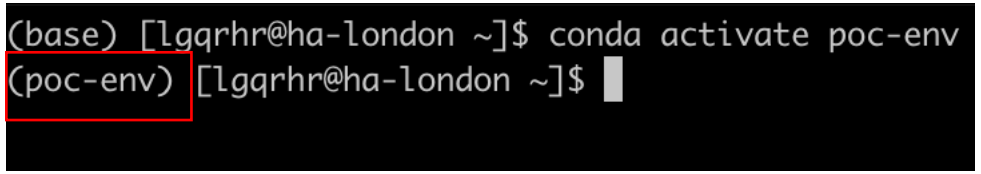
Contents

- ❑ Model Compilation for LPU
- ❑ LPU Execution
  1. HuggingFace API
  2. Serving API
- ❑ Useful commands
- ❑ Troubleshooting

Model Compilation for LPU

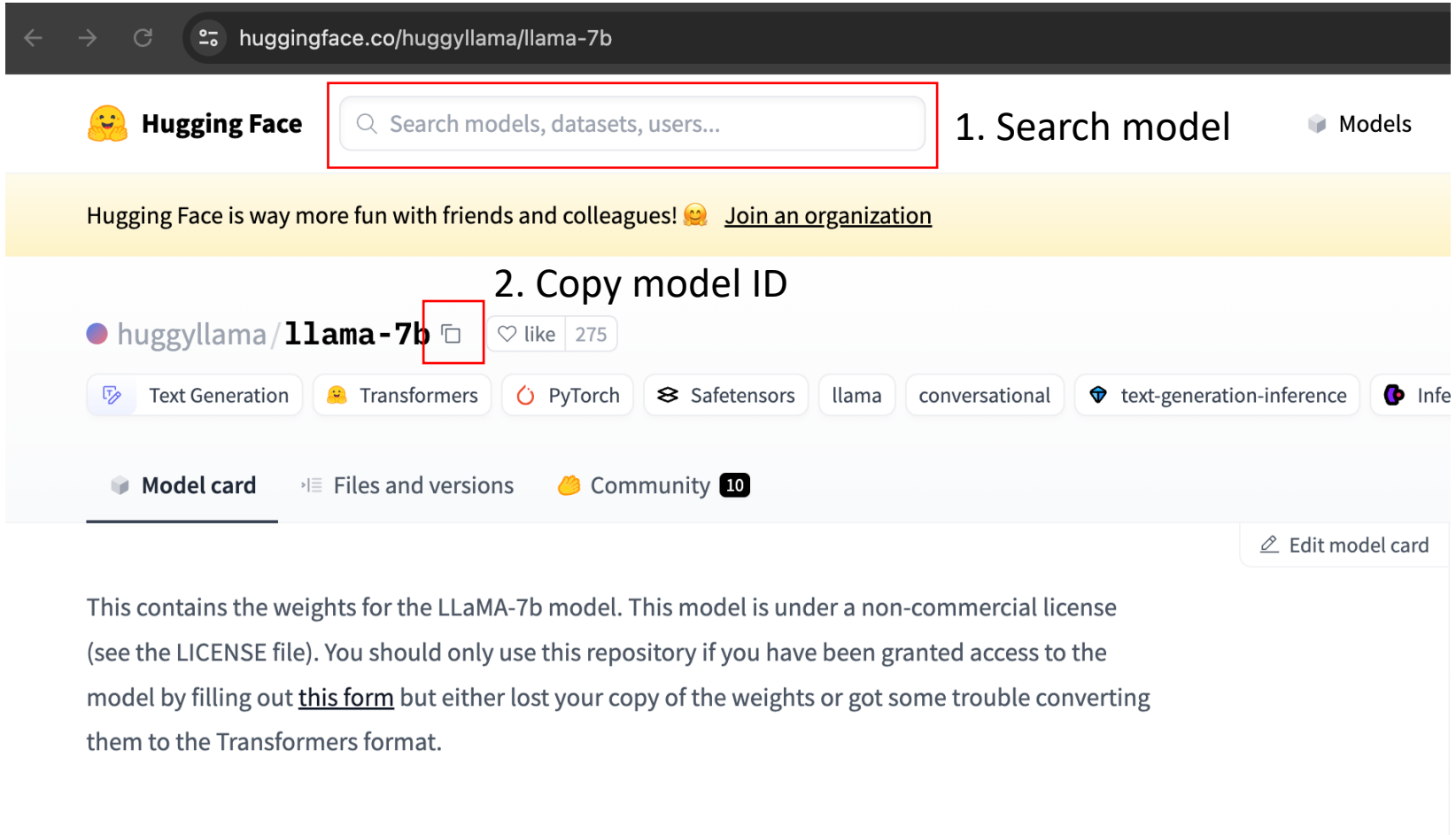
- Before you start, ensure that the python environment is set to “poc-env”

  $conda activate poc-env



Model Compilation for LPU

- Copy “Model ID” in huggingface.co



HYPER ACCEL

Model Compilation for LPU

- Run “hyperdex_sdk/cli/run.py” and paste model ID

1. Run hyperdex_sdk

```
(lpu-env) hyunjun@ha-sydney:~/hyperdex_sdk/cli$ python run.py
**********************************************************************
** HyperDex Model SDK Version 1.3.1
** Copyright (c) 2024 HyperAccel
**********************************************************************

** Phase-1 : Download model from HuggingFace Hub **

In this phase, download the HuggingFace model and save it to the prefix (=/opt/hyperdex/models).
ex) facebook/opt-1.3b ==> download at "/opt/hyperdex/models/facebook/opt-1.3b/ckpt"

Please input the model available on Hugging Face from the options below.
If you are using a custom model, make sure to place it in the prefix path beforehand for it to be usable.

[Option ] Please enter the HuggingFace model id: huggyllama/llama-7b
```

2. Paste model ID

HYPER ACCEL

Model Compilation for LPU

- Set the number of LPU

```
** Phase-2 : Convert the checkpoint format to HyperDex style **

In this phase, we convert the Hugging Face checkpoint into a format that HyperDex can read.
The converter's output includes both a JSON file and a binary file.

[Info   ] Convert the model to HyperDex model format
Loading checkpoint shards: 100%|                                                | 6/6 [00:03<00:00,  1.99it/s]
[Info   ] Save the converted checkpoint at /opt/hyperdex/models/huggyllama/llama-7b/ckpt

** Phase-3 : Generate Optimized Model Mapping **

In this phase, we optimize model mapping for streamlined memory access and efficient model parallelism.
The result of this optimization process is a binary file.

[Option ] Please enter the number of LPU devices:
```

HYPER ACCEL

Model Compilation for LPU

- Done compiling

```
** Phase-3 : Generate Optimized Model Mapping **

In this phase, we optimize model mapping for streamlined memory access and efficient model parallelism.
The result of this optimization process is a binary file.

[Option ] Please enter the number of LPU devices: 1
[Info   ] Optimize the model paramater
[Info   ] Save the optimized data at /opt/hyperdex/models/huggyllama/llama-7b/param

** Phase-4 : Generate Optimized Model Instruction **

In this phase, we analyze the structure of the model to generate optimized instructions.
We use techniques such as Kernel Fusion and Layer Reordering to make hardware processing more efficient.
The result of this optimization is a binary file.
[Info   ] Optimize the model instruction
[Info   ] Save the optimized instruction at /opt/hyperdex/models/huggyllama/llama-7b/inst
[Info   ] Exit SDK program
(lpu-env) hyunjun@ha-sydney:~/hyperdex_sdk/cli$
```

HYPER ACCEL

LPU Execution: HuggingFace API

- Run “examples/text_generation.py”

```python
import sys
#Import hyperdex-python
from hyperdex.transformers import AutoModelForCausalLM, AutoTokenizer

################################################################################
## Main
################################################################################

def main():

  # Hyperdex checkpoint path
  hyperdex_ckpt = "/opt/hyperdex/models/huggyllama/llama-7b"

  # Load model and tokenizer
  model = AutoModelForCausalLM.from_pretrained(hyperdex_ckpt, device_map={"gpu": 0, "lpu": 1})
  tokenizer = AutoTokenizer.from_pretrained(hyperdex_ckpt)

  # Tokenize input context
  inputs = "Hello, my name"
  input_ids = tokenizer.encode(inputs)
```

1. Set model ID and the number of LPUs

HYPER ACCEL

LPU Execution: HuggingFace API

- Generated text will be displayed

```
(lpu-env) hyunjun@ha-sydney:~/examples$ python text_generation.py
Hello, my name is [Your Name] and I am a [Your Position] at [Your Company].
 I am writing to express my interest in the [Job Title] position
```

LPU Execution: Serving API

- Run “examples/launch.sh” to open the serving system

```
(poc-env) server@mymeta:~/examples/api_server$ ./launch.sh
/home/server/documents/env/envs/poc-env/lib/python3.9/site-packages/torch/_utils
 be removed in the future and UntypedStorage will be the only storage class. Thi
ectly.  To access UntypedStorage directly, use tensor.untyped_storage() instead
  return self.fget.__get__(instance, owner)()
INFO:     Started server process [7198]
INFO:     Waiting for application startup.
INFO:     Application startup complete.
INFO:     Uvicorn running on http://0.0.0.0:8000 (Press CTRL+C to quit)
```

LPU Execution: Serving API

- Open another terminal to send the request
- Generated text will be displayed

```
INFO:     Started server process [25924]
INFO:     Waiting for application startup.
INFO:     Application startup complete.
INFO:     Uvicorn running on http://0.0.0.0:8000 (Press CTRL+C to quit)
INFO:     127.0.0.1:52970 - "POST / HTTP/1.1" 200 OK
INFO:     127.0.0.1:56568 - "POST / HTTP/1.1" 200 OK
```

Server Terminal

```
(poc-env) server@mymeta:~/examples/api_server$ python client.py
Hello, my name is John. I am a professional writer and editor who has been writing for over
 20 years. My work includes articles on business, finance, real estate, health care
```

User Terminal

HYPER ACCEL

Useful commands: hyperdex-smi

- “hyperdex-smi” shows status of the LPUs (same as nvidia-smi)

```
(lpu-env) hyunjun@ha-sydney:~/examples$ hyperdex-smi
Fri Jul 26 16:59:26 2024
+-----------------------------------------------------------------------------+
| HYPERDEX-SMI            XRT Version: 2022.2     HyperDex Version: 1.3.2     |
+-------------------------------+----------------------+----------------------+
| FPGA Name        Persistence-M| Bus-Id     Bitstream | Volatile Uncorr. ECC |
|      Temp  Perf  Pwr:Usage/Cap|         Memory-Usage | FPGA-Util Compute M. |
|                               |                      |               MIG M. |
|===============================+======================+======================|
|    0      XILINX U55C    Off  | 0000:cb:00.1      On |                  N/A |
|       46C    P8    33W / 225W |   2880MiB / 16384MiB |     17%      Default |
|                               |                      |                  N/A |
+-------------------------------+----------------------+----------------------+
|    1      XILINX U55C    Off  | 0000:ca:00.1     Off |                  N/A |
|       34C    P8    16W / 225W |      0MiB /     0MiB |      0%      Default |
|                               |                      |                  N/A |
+-------------------------------+----------------------+----------------------+
|    2      XILINX U55C    Off  | 0000:b2:00.1     Off |                  N/A |
|       34C    P8    17W / 225W |      0MiB /     0MiB |      0%      Default |
|                               |                      |                  N/A |
+-------------------------------+----------------------+----------------------+
|    3      XILINX U55C    Off  | 0000:b1:00.1     Off |                  N/A |
|       34C    P8    18W / 225W |      0MiB /     0MiB |      0%      Default |
|                               |                      |                  N/A |
+-------------------------------+----------------------+----------------------+
```

HYPER ACCEL

Useful commands: hyperdex-reset

- If the LPUs are not functioning properly, “hyperdex-reset” initializes LPUs

```
(lpu-env) hyunjun@ha-sydney:~/examples$ hyperdex-reset
Reset LPU-0 card complete!
Reset LPU-1 card complete!
Reset LPU-2 card complete!
Reset LPU-3 card complete!
Reset LPU-4 card complete!
Reset LPU-5 card complete!
```

Useful commands: xbuil examine

- "xbutil examine" verifies if the LPU is installed properly

```
(lpu-env) hyunjun@ha-sydney:~/examples$ xbutil examine
System Configuration
  OS Name              : Linux
  Release              : 5.15.0-25-generic
  Version              : #25-Ubuntu SMP Wed Mar 30 15:54:22 UTC 2022
  Machine              : x86_64
  CPU Cores            : 48
  Memory               : 257574 MB
  Distribution         : Ubuntu 22.04.4 LTS
  GLIBC                : 2.35
  Model                : ESC4000-E10

XRT
  Version              : 2.14.354
  Branch               : 2022.2
  Hash                 : 43926231f7183688add2dccfd391b36a1f000bea
  Hash Date            : 2022-10-08 09:49:53
  XOCL                 : 2.14.354, 43926231f7183688add2dccfd391b36a1f000bea
  XCLMGMT              : 2.14.354, 43926231f7183688add2dccfd391b36a1f000bea

Devices present
BDF             :  Shell                          Platform UUID                        Device ID       Device Ready*
-------------------------------------------------------------------------------------------------------------------
[0000:cb:00.1]  :  xilinx_u55c_gen3x16_xdma_base_3  97088961-FEAE-DA91-52A2-1D9DFD63CCEF  user(inst=129)  Yes
[0000:ca:00.1]  :  xilinx_u55c_gen3x16_xdma_base_3  97088961-FEAE-DA91-52A2-1D9DFD63CCEF  user(inst=128)  Yes
[0000:b2:00.1]  :  xilinx_u55c_gen3x16_xdma_base_3  97088961-FEAE-DA91-52A2-1D9DFD63CCEF  user(inst=133)  Yes
[0000:b1:00.1]  :  xilinx_u55c_gen3x16_xdma_base_3  97088961-FEAE-DA91-52A2-1D9DFD63CCEF  user(inst=132)  Yes
[0000:32:00.1]  :  xilinx_u55c_gen3x16_xdma_base_3  97088961-FEAE-DA91-52A2-1D9DFD63CCEF  user(inst=131)  Yes
[0000:31:00.1]  :  xilinx_u55c_gen3x16_xdma_base_3  97088961-FEAE-DA91-52A2-1D9DFD63CCEF  user(inst=130)  Yes
```

Useful commands: lspci | grep Xilinx

- "lspci | grep Xilinx" checks if the device is physically recognized in the PCIe slot

```
(lpu-env) hyunjun@ha-sydney:~/examples$ lspci | grep Xilinx
31:00.0 Processing accelerators: Xilinx Corporation Device 505c
31:00.1 Processing accelerators: Xilinx Corporation Device 505d
32:00.0 Processing accelerators: Xilinx Corporation Device 505c
32:00.1 Processing accelerators: Xilinx Corporation Device 505d
b1:00.0 Processing accelerators: Xilinx Corporation Device 505c
b1:00.1 Processing accelerators: Xilinx Corporation Device 505d
b2:00.0 Processing accelerators: Xilinx Corporation Device 505c
b2:00.1 Processing accelerators: Xilinx Corporation Device 505d
ca:00.0 Processing accelerators: Xilinx Corporation Device 505c
ca:00.1 Processing accelerators: Xilinx Corporation Device 505d
cb:00.0 Processing accelerators: Xilinx Corporation Device 505c
cb:00.1 Processing accelerators: Xilinx Corporation Device 505d
```

Troubleshooting

❑Problem

1. [XRT] ERROR: Kernel arg 'axi00_ptr0' is not set

```
[XRT] ERROR: Operation failed due to earlier error 'std::bad_alloc'
Return code for cl::enqueueReadBuffer flags=-5
Hello, my name
[XRT] ERROR: Kernel arg 'axi00_ptr0' is not set
[XRT] ERROR: Kernel arg 'axi00_ptr0' is not set
[XRT] ERROR: Kernel arg 'axi00_ptr0' is not set
[XRT] ERROR: Kernel arg 'axi00_ptr0' is not set
```

❑Solution

1. If the computer is rebooted, you should enable the host memory.
2. Run the “**document/scripts/host_memory_enable.sh**”

Troubleshooting

❑Problem

1. [XRT] ERROR: unable to sync BO: Input/output error
2. Deadrock situation

```
Build hash: 43926231f7183688add2dccfd391b36a1f000bea
Build date: 2022-10-08 09:49:53
Git branch: 2022.2
PID: 146196
UID: 2007
[Fri Jul 26 08:16:58 2024 GMT]
HOST: ha-sydney
EXE: /home/members/hyunjun/miniconda3/envs/lpu-env/bin/python3.11
[XRT] ERROR: unable to sync BO: Input/output error
terminate called after throwing an instance of 'xrt_xocl::error'
  what():  event 138 never submitted
```

❑Solution

1. This error occurs when unintended signals are sent to the device such as ctrl+C
2. Run “**hyperdex-reset**” and then execute it again.

More information?

- Our website: https://hyperaccel.ai
- Our docs: https://docs.hyperaccel.ai
- Linkedin: https://linkedin.com/company/hyperaccel
- Contact us: contact@hyperaccel.ai

HYPER ACCEL