

```
azam@polaris.alcf.anl.gov x ~ /cafs/richie x + v x
pip list | grep pandas
numpy 1.21.6
pandas 1.3.5
essing/bert$ python bert_tiny.py/groqflow/proof_points/natural_language_proce
tokenizer_config.json: 100%| 346/346 [00:00<00:00, 3.99MB/s]
vocab.txt: 100%| 232k/232k [00:00<00:00, 4.93MB/s]
special_tokens_map.json: 100%| 112/112 [00:00<00:00, 1.68MB/s]
config.json: 100%| 760/760 [00:00<00:00, 6.15MB/s]
pytorch_model.bin: 100%| 17.6M/17.6M [00:00<00:00, 289MB/s]
/home/azam/miniconda3/envs/groqflow/lib/python3.10/site-packages/torch/_utils.py:831: UserWarning: TypedStorage is deprecated. It will be removed in the fut
ure and UntypedStorage will be the only storage class. This should only matter to you if you are using storages directly. To access UntypedStorage directly
, use tensor.untyped_storage() instead of tensor.storage()
return self.fget.__get__(instance, owner)()

Building "bert_tiny"
✓ Exporting PyTorch to ONNX
✓ Optimizing ONNX file
✓ Checking for Op support
✓ Converting to FP16
✓ Compiling model
✓ Assembling model

Woohoo! Saved to ~/.cache/groqflow/bert_tiny
Preprocessing data.
/home/azam/miniconda3/envs/groqflow/lib/python3.10/site-packages/datasets/load.py:1461: FutureWarning: The repository for sst contains custom code which mus
t be executed to correctly load the dataset. You can inspect the repository content at https://hf.co/datasets/sst
You can avoid this message in future by passing the argument 'trust_remote_code=True'.
Passing 'trust_remote_code=True' will be mandatory to load this dataset from the next major release of 'datasets'.
warnings.warn(
Downloading builder script: 100%| 9.13k/9.13k [00:00<00:00, 47.9MB/s]
Downloading readme: 100%| 6.68k/6.68k [00:00<00:00, 50.4MB/s]
Downloading data: 100%| 6.37M/6.37M [00:01<00:00, 5.98MB/s]
Downloading data: 100%| 790k/790k [00:00<00:00, 1.46MB/s]
Generating train split: 100%| 8544/8544 [00:00<00:00, 11423.52 examples/s]
Generating validation split: 100%| 1101/1101 [00:00<00:00, 2036.46 example
Generating test split: 100%| 2210/2210 [00:00<00:00, 3963.79 examples/s]

Info: No inputs received for benchmark. Using the inputs provided during model compilation.
Running inference on GroqChip.
```



```
azam@polaris.alcf.anl.gov  x  ~/cafs/riche  x  +  v  -  [  ]  x
ure and UntypedStorage will be the only storage class. This should only matter to you if you are using storages directly. To access UntypedStorage directly
, use tensor.untyped_storage() instead of tensor.storage()
return self.fget.__get__(instance, owner)()

Building "bert_tiny"
✓ Exporting PyTorch to ONNX
✓ Optimizing ONNX file
✓ Checking for Op support
✓ Converting to FP16
✓ Compiling model
✓ Assembling model

Woohoo! Saved to ~/.cache/groqflow/bert_tiny
Preprocessing data.
/home/azam/miniconda3/envs/groqflow/lib/python3.10/site-packages/datasets/load.py:1461: FutureWarning: The repository for sst contains custom code which mus
t be executed to correctly load the dataset. You can inspect the repository content at https://hf.co/datasets/sst
You can avoid this message in future by passing the argument `trust_remote_code=True`.
Passing `trust_remote_code=True` will be mandatory to load this dataset from the next major release of `datasets`.
warnings.warn(
Downloading builder script: 100%|██████████| 9.13k/9.13k [00:00<00:00, 47.9MB/s]
Downloading readme: 100%|██████████| 6.68k/6.68k [00:00<00:00, 50.4MB/s]
Downloading data: 100%|██████████| 6.37M/6.37M [00:01<00:00, 5.98MB/s]
Downloading data: 100%|██████████| 790k/790k [00:00<00:00, 1.46MB/s]
Generating train split: 100%|██████████| 8544/8544 [00:00<00:00, 11423.52 examples/s]
Generating validation split: 100%|██████████| 1101/1101 [00:00<00:00, 2036.46 example
Generating test split: 100%|██████████| 2210/2210 [00:00<00:00, 3963.79 examples/s]

Info: No inputs received for benchmark. Using the inputs provided during model compilation.
Running inference on GroqChip.
Running inference using PyTorch model (CPU).
100%|██████████| 2210/2210 [00:04<00:00, 442.99it/s]
+-----+-----+-----+-----+-----+-----+
| Source | Accuracy | end-to-end latency (ms) | end-to-end IPS | on-chip latency (ms) | on-chip IPS |
+-----+-----+-----+-----+-----+-----+
| cpu    | 77.47%   | 2.26                   | 442.94         | --                 | --          |
| groq   | 77.47%   | 0.05                   | 18331.84       | 0.03               | 37576.72    |
+-----+-----+-----+-----+-----+-----+

Proof point /home/azam/groqflow/proof_points/natural_language_processing/bert/bert_tiny.py finished!
essing/bert$ am@groq-r01-gn-08:~/groqflow/proof_points/natural_language_proce
```