Profiling Transformer-Based Model

Wannaphong Phattiyaphaibun 2130810
Patomporn Payoungkhamdee 2130807

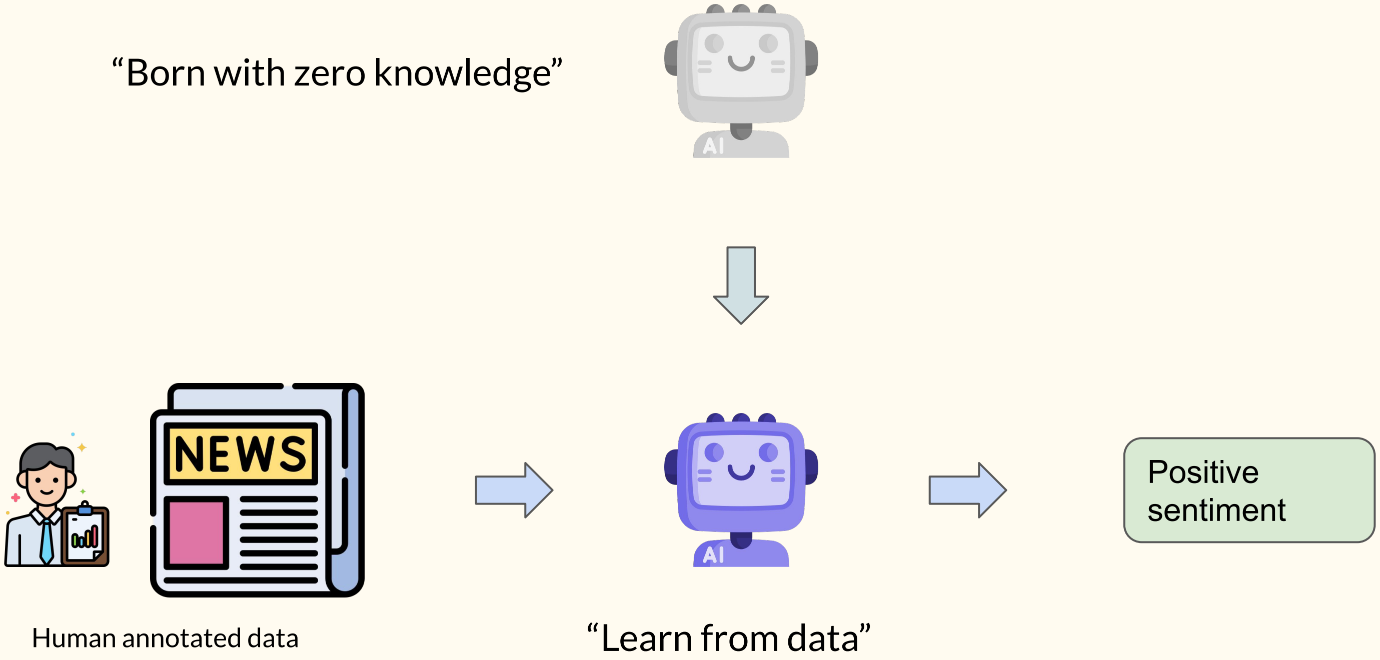
ComArch Project
IST 503, VISTEC

Outline

- Background
- Motivation
- Methodology
- Profiling
  - BERT
  - GPT
- Summary

Background

Transfer Learning



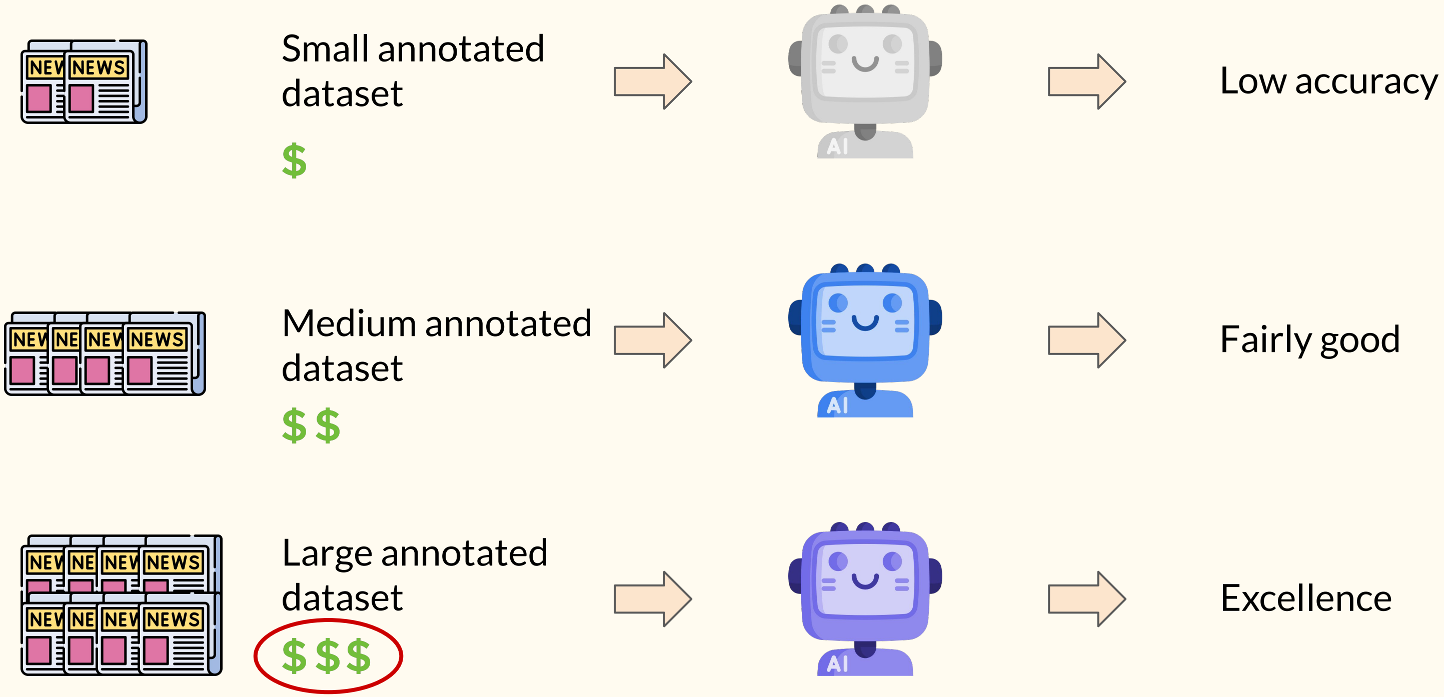
“Born with zero knowledge”

Human annotated data

“Learn from data”

Positive sentiment

Deep Learning (Not Transfer Learning)



Small annotated dataset

$

Low accuracy

Medium annotated dataset

$ $

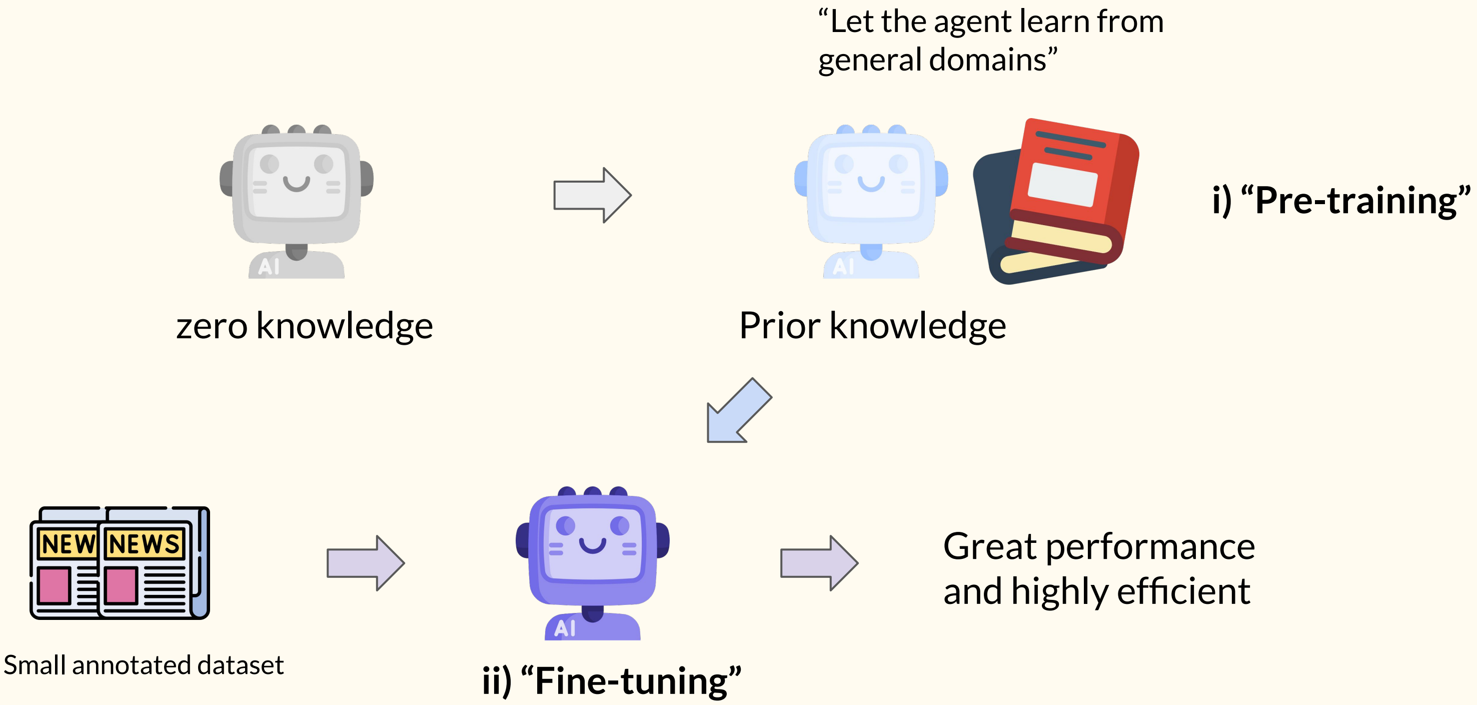
Fairly good

Large annotated dataset

$ $ $

Excellence

Transfer Learning



"Let the agent learn from general domains"

zero knowledge

Prior knowledge

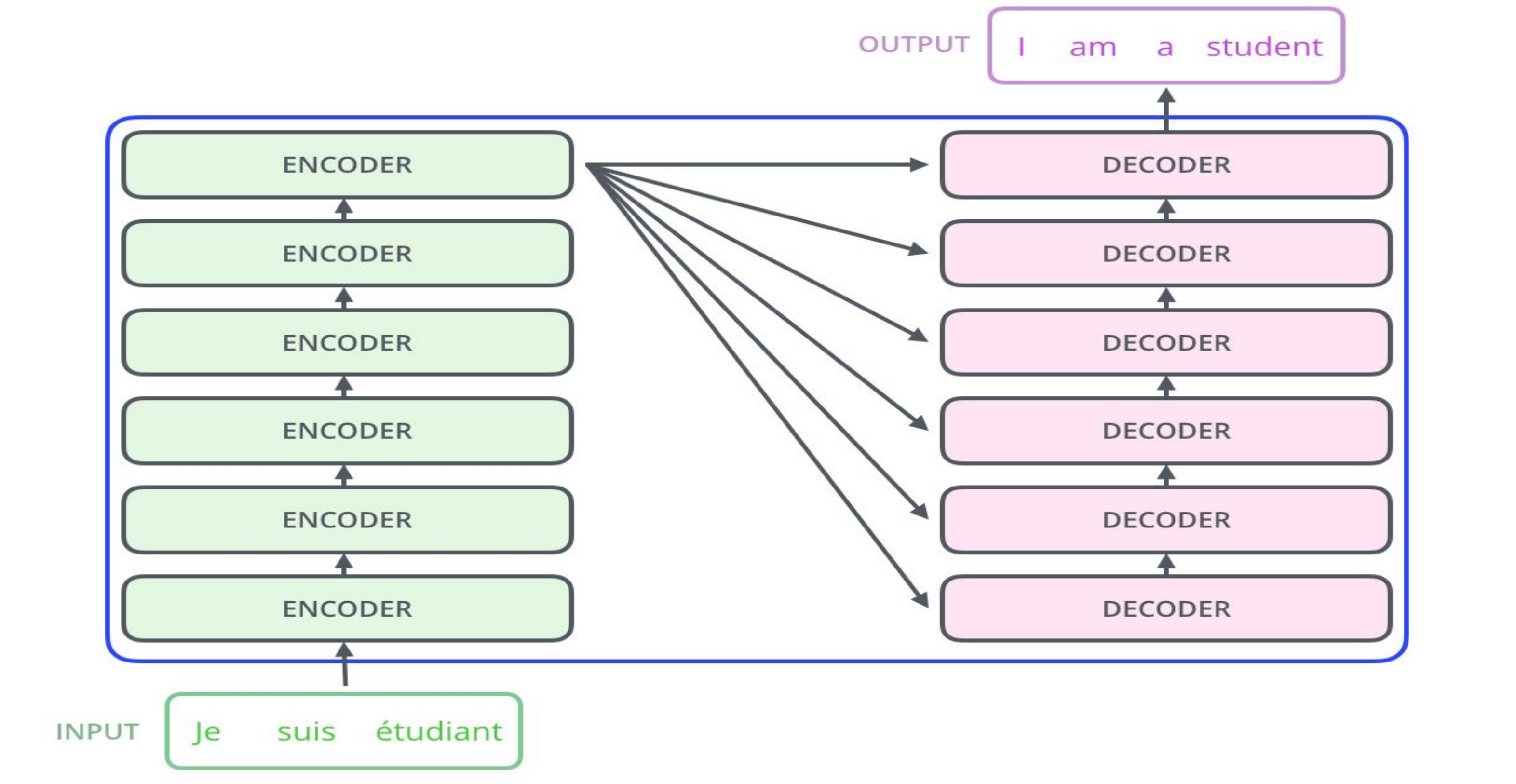
i) "Pre-training"

Small annotated dataset

ii) "Fine-tuning"

Great performance and highly efficient

Transformers



The Illustrated Transformer by Jay Alammar https://jalammar.github.io/illustrated-transformer/

Transformers

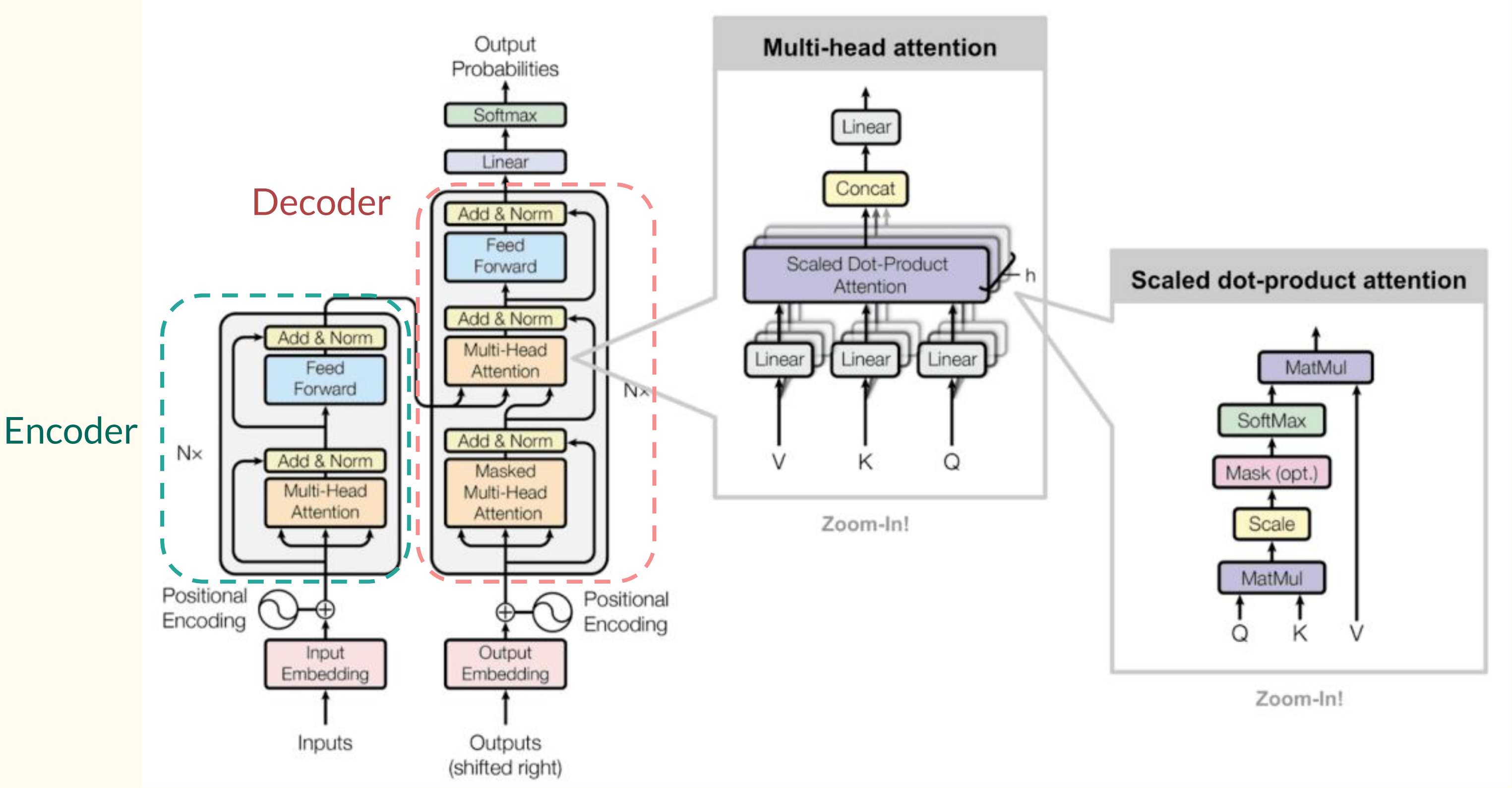
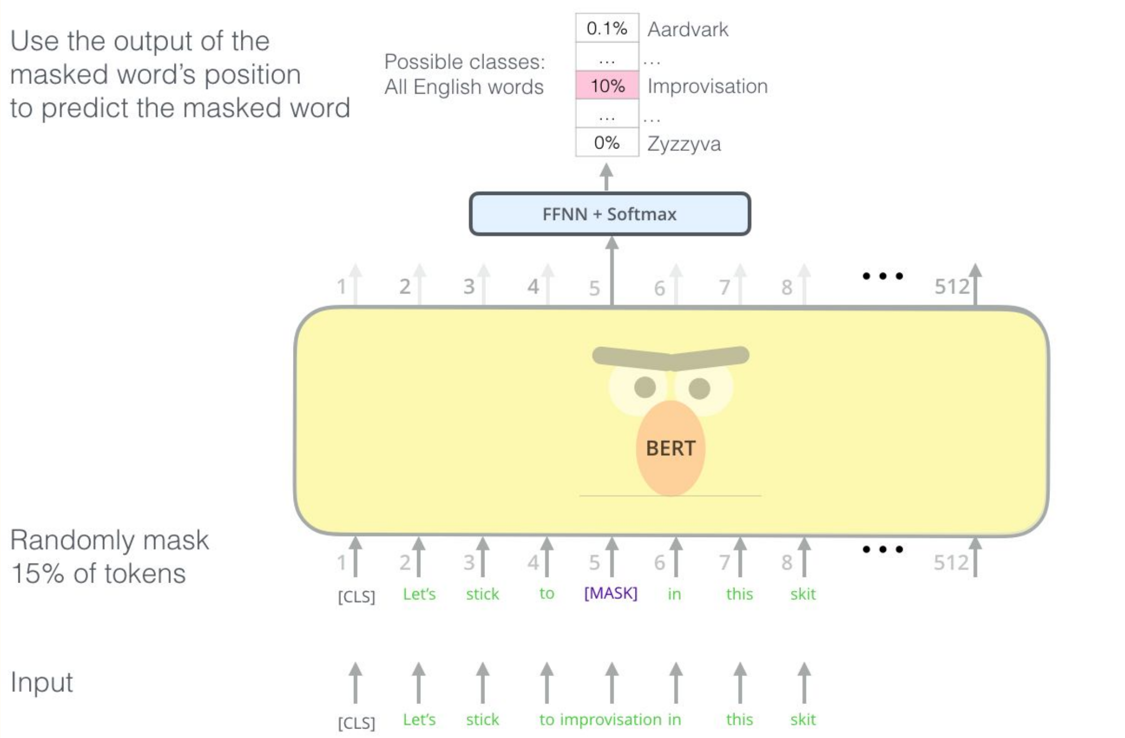


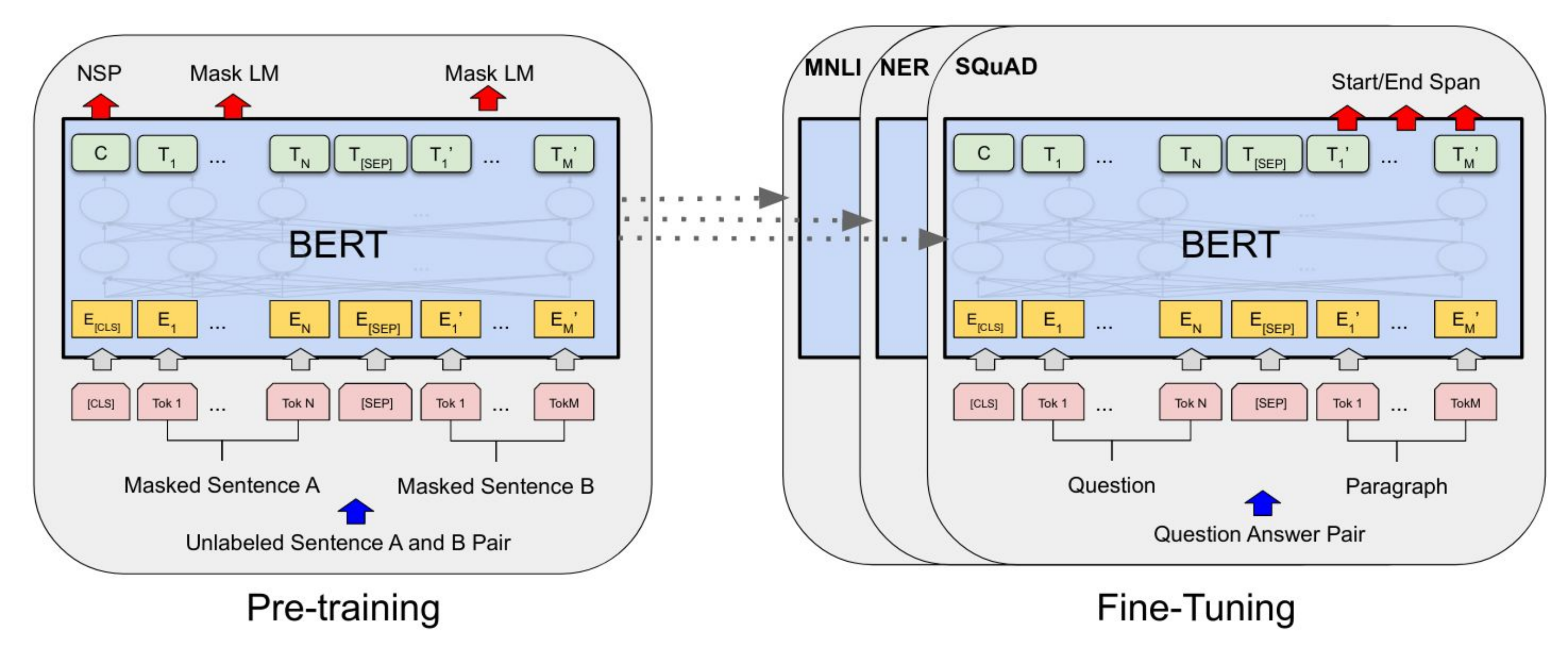
Image taken from https://neptune.ai/blog/natural-language-processing-with-hugging-face-and-transformers

BERT: Tweaking the encoder part


Use the output of the masked word's position to predict the masked word
Possible classes: All English words
0.1% Aardvark
... ...
10% Improvisation
... ...
0% Zyzzyva
FFNN + Softmax
1 2 3 4 5 6 7 8 ... 512
BERT
1 2 3 4 5 6 7 8 ... 512
Randomly mask 15% of tokens
[CLS] Let's stick to [MASK] in this skit
Input
[CLS] Let's stick to improvisation in this skit

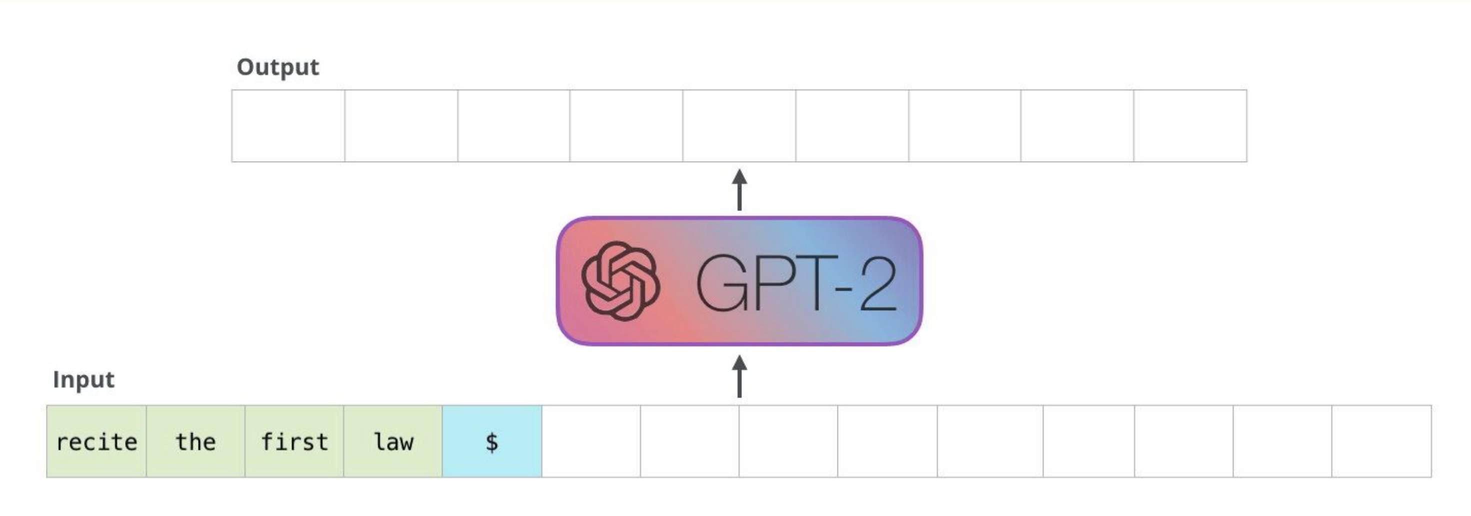
The Illustrated BERT, ELMo, and co. (How NLP Cracked Transfer Learning) by Jay Alammar https://jalammar.github.io/illustrated-bert/

BERT: Tweaking the encoder part



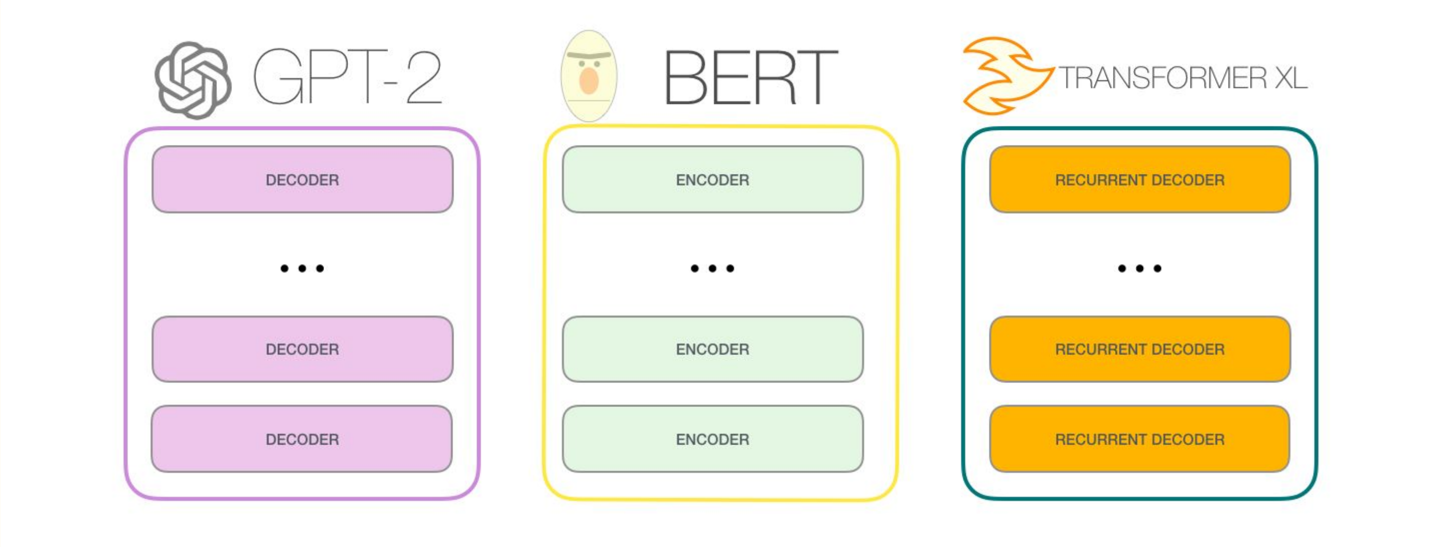
[Devlin et. al (2019), URL: https://aclanthology.org/N19-1423.pdf]

GPT-2

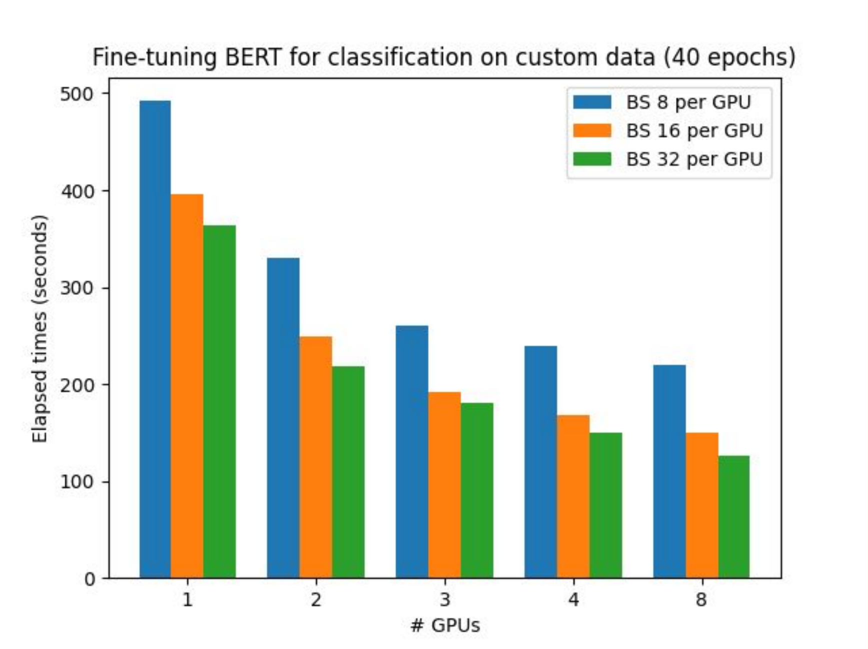


The Illustrated GPT-2 (Visualizing Transformer Language Models) by Jay Alammar
https://jalammar.github.io/illustrated-gpt2/

GPT-2



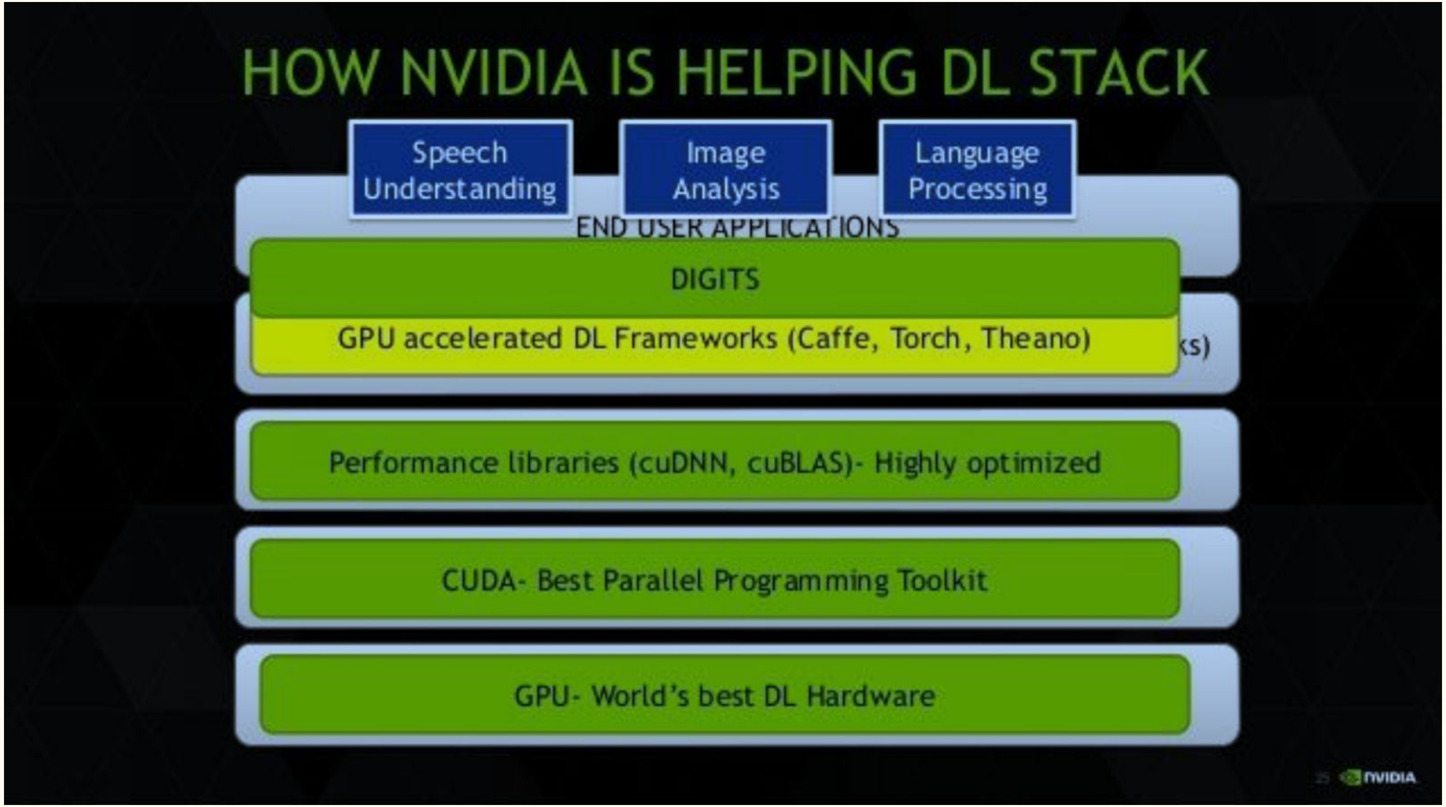
The Illustrated GPT-2 (Visualizing Transformer Language Models) by Jay Alammar
https://jalammar.github.io/illustrated-gpt2/


Fine-tuning BERT for classification on custom data (40 epochs)
BS 8 per GPU
BS 16 per GPU
BS 32 per GPU
Elapsed times (seconds)
0
100
200
300
400
500
1
2
3
4
8
# GPUs

Motivation

Methodology

What happen under the hood



HOW NVIDIA IS HELPING DL STACK

Speech Understanding

Image Analysis

Language Processing

END USER APPLICATIONS

DIGITS

GPU accelerated DL Frameworks (Caffe, Torch, Theano)

ks)

Performance libraries (cuDNN, cuBLAS)- Highly optimized

CUDA- Best Parallel Programming Toolkit

GPU- World's best DL Hardware

Schematic of a GPU

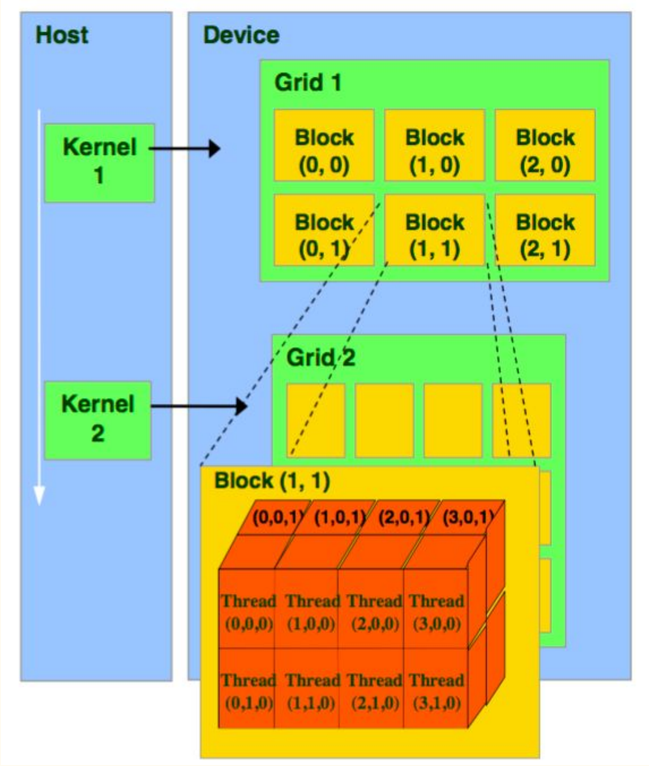


Image taken from https://nyu-cds.github.io/python-gpu/02-cuda/

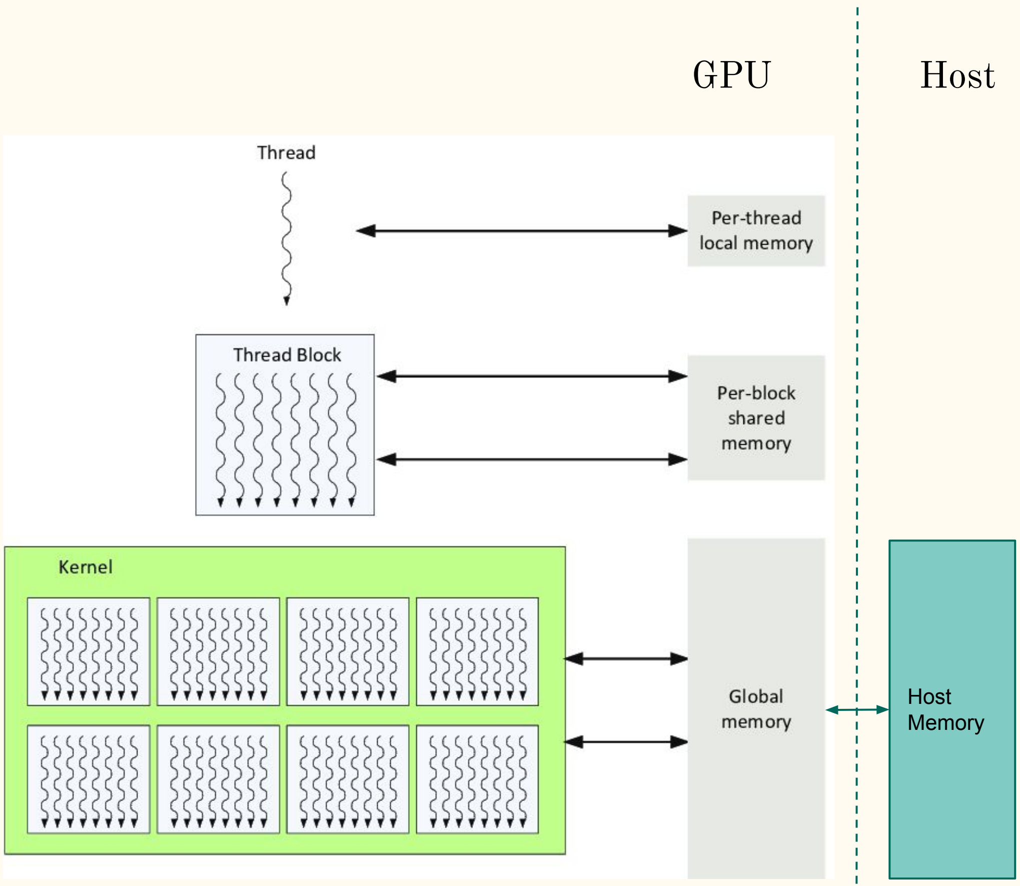
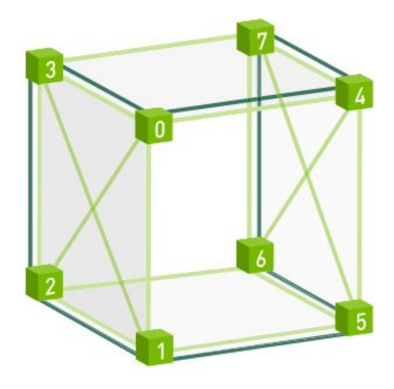


Image is modified from 10.1587/elex.14.20170373

DGX-1
V100
Network Topology



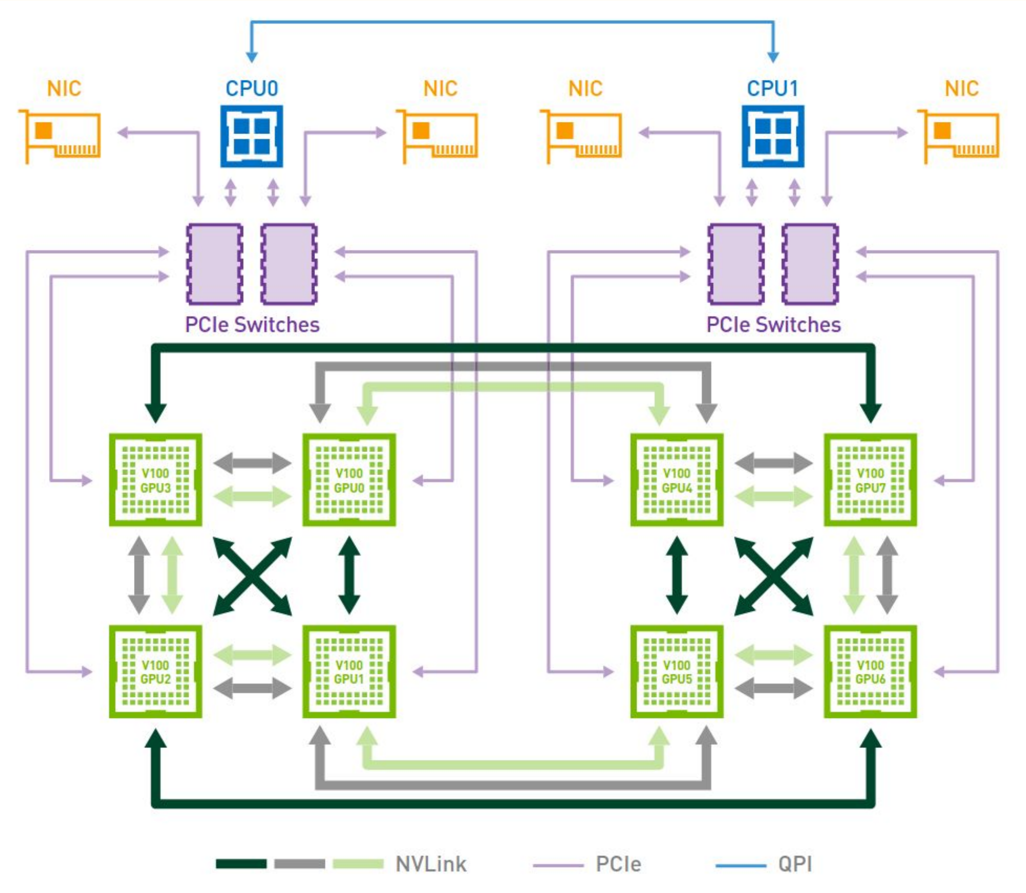
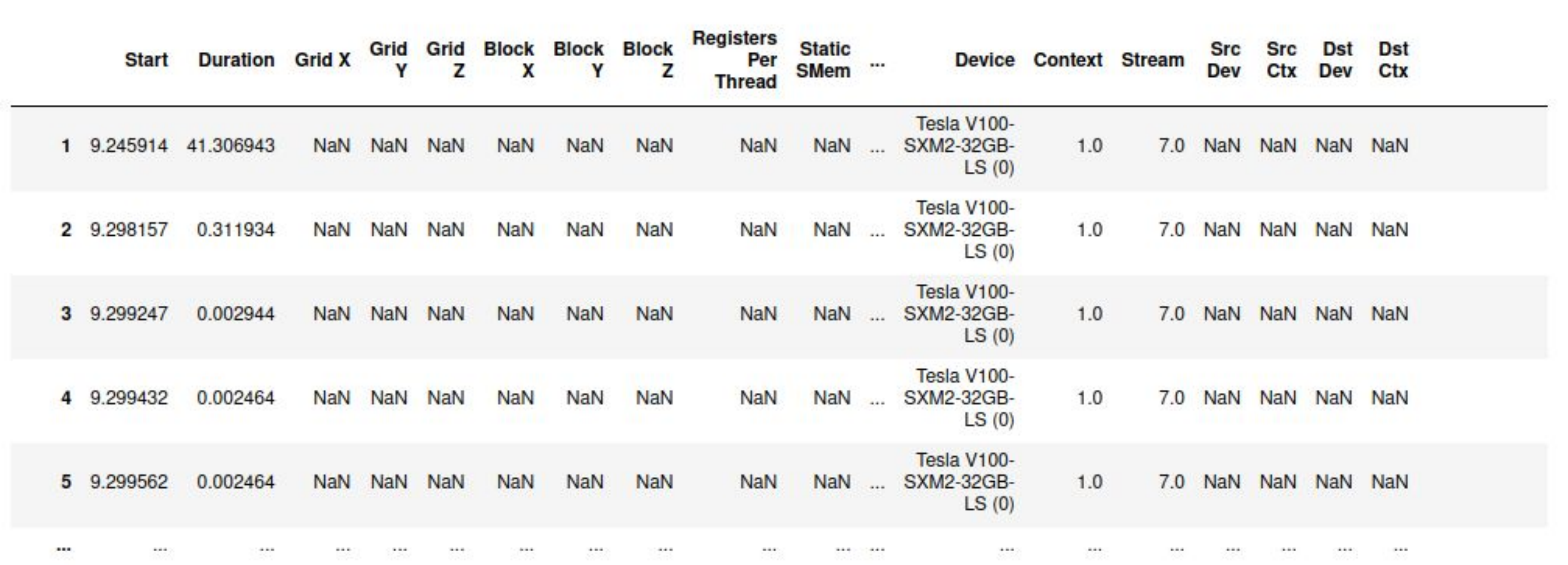


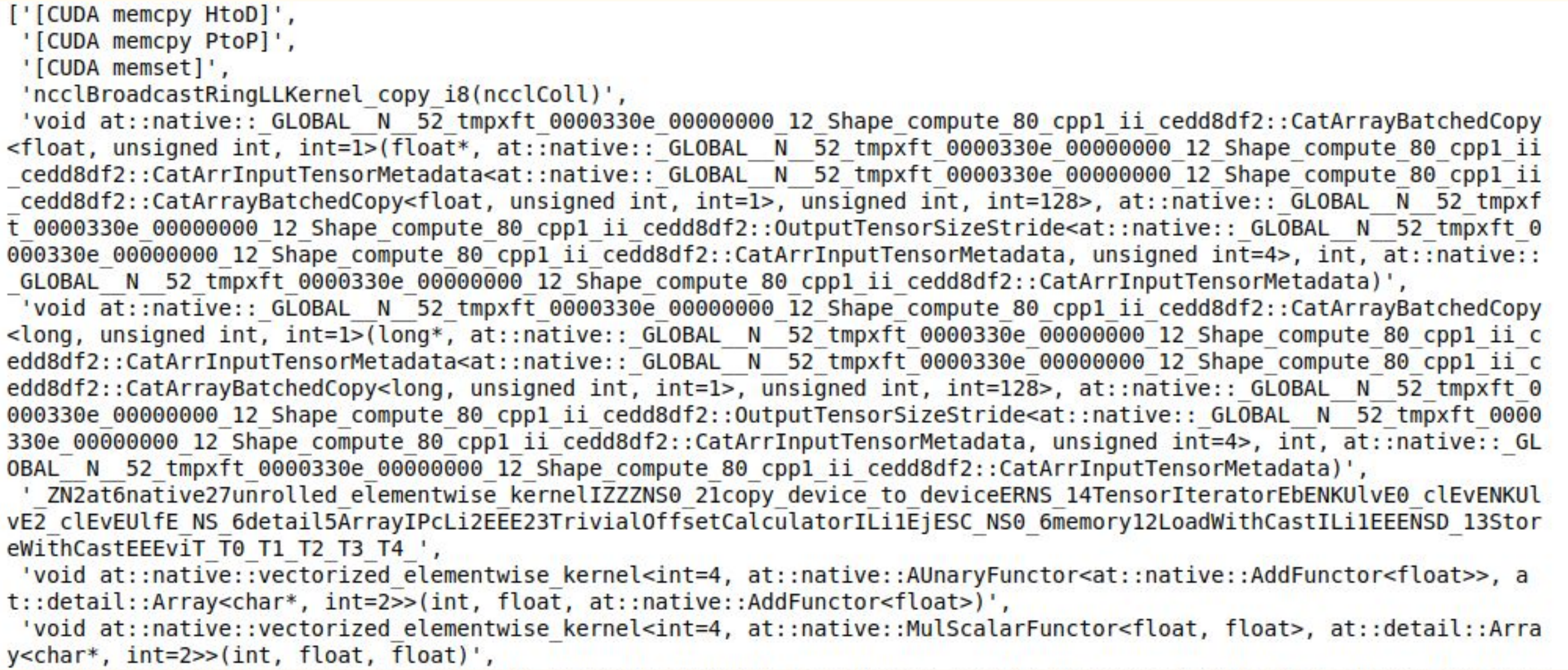
Image taken from https://images.nvidia.com/content/pdf/dgx1-v100-system-architecture-whitepaper.pdf

Raw Profiling Output



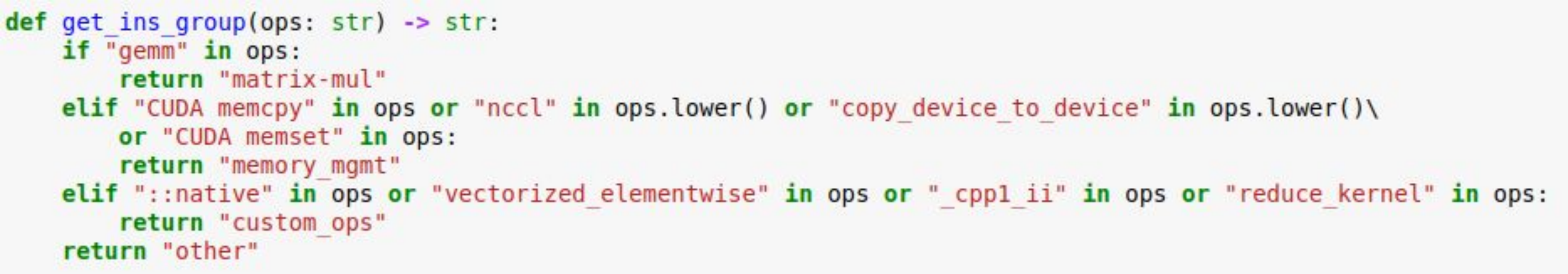
	Start	Duration	Grid X	Grid Y	Grid Z	Block X	Block Y	Block Z	Registers Per Thread	Static SMem	...	Device	Context	Stream	Src Dev	Src Ctx	Dst Dev	Dst Ctx
1	9.245914	41.306943	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	Tesla V100-SXM2-32GB-LS (0)	1.0	7.0	NaN	NaN	NaN	NaN
2	9.298157	0.311934	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	Tesla V100-SXM2-32GB-LS (0)	1.0	7.0	NaN	NaN	NaN	NaN
3	9.299247	0.002944	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	Tesla V100-SXM2-32GB-LS (0)	1.0	7.0	NaN	NaN	NaN	NaN
4	9.299432	0.002464	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	Tesla V100-SXM2-32GB-LS (0)	1.0	7.0	NaN	NaN	NaN	NaN
5	9.299562	0.002464	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	Tesla V100-SXM2-32GB-LS (0)	1.0	7.0	NaN	NaN	NaN	NaN
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...

Some of Collected Interface Names



```
['[CUDA memcpy HtoD]',
 '[CUDA memcpy PtoP]',
 '[CUDA memset]',
 'ncclBroadcastRingLLKernel_copy_i8(ncclColl)',
 'void at::native::_GLOBAL__N__52_tmpxft_0000330e_00000000_12_Shape_compute_80_cpp1_ii_cedd8df2::CatArrayBatchedCopy
<float, unsigned int, int=1>(float*, at::native::_GLOBAL__N__52_tmpxft_0000330e_00000000_12_Shape_compute_80_cpp1_ii
_cedd8df2::CatArrInputTensorMetadata<at::native::_GLOBAL__N__52_tmpxft_0000330e_00000000_12_Shape_compute_80_cpp1_ii
_cedd8df2::CatArrayBatchedCopy<float, unsigned int, int=1>, unsigned int, int=128>, at::native::_GLOBAL__N__52_tmpxf
t_0000330e_00000000_12_Shape_compute_80_cpp1_ii_cedd8df2::OutputTensorSizeStride<at::native::_GLOBAL__N__52_tmpxft_0
000330e_00000000_12_Shape_compute_80_cpp1_ii_cedd8df2::CatArrInputTensorMetadata, unsigned int=4>, int, at::native::
_GLOBAL__N__52_tmpxft_0000330e_00000000_12_Shape_compute_80_cpp1_ii_cedd8df2::CatArrInputTensorMetadata)',
 'void at::native::_GLOBAL__N__52_tmpxft_0000330e_00000000_12_Shape_compute_80_cpp1_ii_cedd8df2::CatArrayBatchedCopy
<long, unsigned int, int=1>(long*, at::native::_GLOBAL__N__52_tmpxft_0000330e_00000000_12_Shape_compute_80_cpp1_ii_c
edd8df2::CatArrInputTensorMetadata<at::native::_GLOBAL__N__52_tmpxft_0000330e_00000000_12_Shape_compute_80_cpp1_ii_c
edd8df2::CatArrayBatchedCopy<long, unsigned int, int=1>, unsigned int, int=128>, at::native::_GLOBAL__N__52_tmpxft_0
000330e_00000000_12_Shape_compute_80_cpp1_ii_cedd8df2::OutputTensorSizeStride<at::native::_GLOBAL__N__52_tmpxft_0000
330e_00000000_12_Shape_compute_80_cpp1_ii_cedd8df2::CatArrInputTensorMetadata, unsigned int=4>, int, at::native::_GL
OBAL__N__52_tmpxft_0000330e_00000000_12_Shape_compute_80_cpp1_ii_cedd8df2::CatArrInputTensorMetadata)',
 '_ZN2at6native27unrolled_elementwise_kernelIZZZNS0_21copy_device_to_deviceERNS_14TensorIteratorEbENKUlvE0_clEvENKUl
vE2_clEvEUlfE_NS_6detail5ArrayIPcLi2EEE23TrivialOffsetCalculatorILi1EjESC_NS0_6memory12LoadWithCastILi1EEENSD_13Stor
eWithCastEEEviT_T0_T1_T2_T3_T4_',
 'void at::native::vectorized_elementwise_kernel<int=4, at::native::AUnaryFunctor<at::native::AddFunctor<float>>, a
t::detail::Array<char*, int=2>>(int, float, at::native::AddFunctor<float>)',
 'void at::native::vectorized_elementwise_kernel<int=4, at::native::MulScalarFunctor<float, float>, at::detail::Arra
y<char*, int=2>>(int, float, float)',
```

Grouping



```python
def get_ins_group(ops: str) -> str:
    if "gemm" in ops:
        return "matrix-mul"
    elif "CUDA memcpy" in ops or "nccl" in ops.lower() or "copy_device_to_device" in ops.lower()\
        or "CUDA memset" in ops:
        return "memory_mgmt"
    elif "::native" in ops or "vectorized_elementwise" in ops or "_cpp1_ii" in ops or "reduce_kernel" in ops:
        return "custom_ops"
    return "other"
```

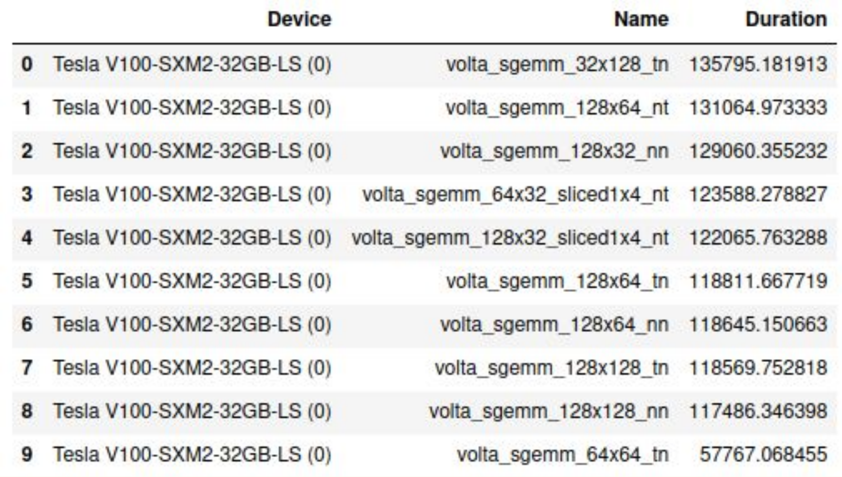
Profiling BERT

Configurations

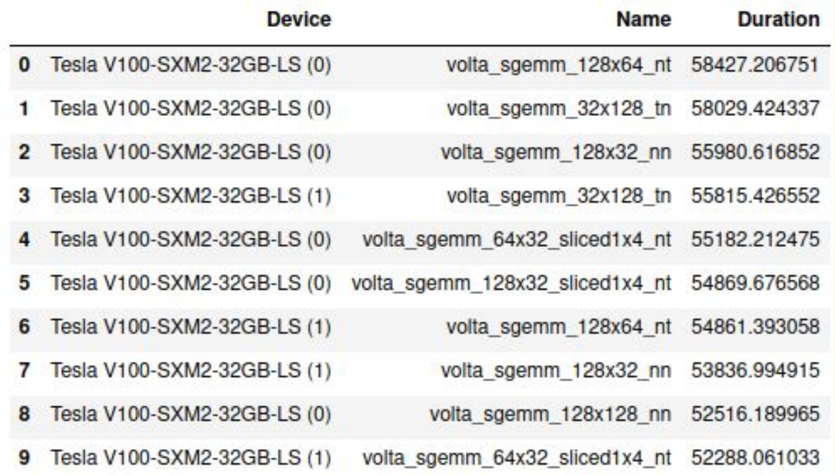
- Batch size: 8
- GPU: 1, 2 and 4
- Max length of Model: 416
- Model name: Wangchanberta (Thai RoBERTa model)
- Dataset: wongnai_reviews (Text classification)
- Epoch: 1

BERT: Top 10

1 GPU

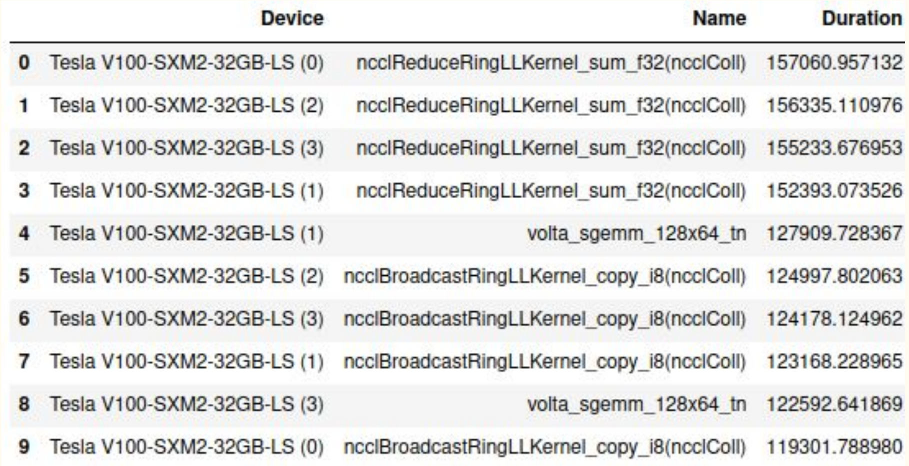


	Device	Name	Duration
0	Tesla V100-SXM2-32GB-LS (0)	volta_sgemm_32x128_tn	135795.181913
1	Tesla V100-SXM2-32GB-LS (0)	volta_sgemm_128x64_nt	131064.973333
2	Tesla V100-SXM2-32GB-LS (0)	volta_sgemm_128x32_nn	129060.355232
3	Tesla V100-SXM2-32GB-LS (0)	volta_sgemm_64x32_sliced1x4_nt	123588.278827
4	Tesla V100-SXM2-32GB-LS (0)	volta_sgemm_128x32_sliced1x4_nt	122065.763288
5	Tesla V100-SXM2-32GB-LS (0)	volta_sgemm_128x64_tn	118811.667719
6	Tesla V100-SXM2-32GB-LS (0)	volta_sgemm_128x64_nn	118645.150663
7	Tesla V100-SXM2-32GB-LS (0)	volta_sgemm_128x128_tn	118569.752818
8	Tesla V100-SXM2-32GB-LS (0)	volta_sgemm_128x128_nn	117486.346398
9	Tesla V100-SXM2-32GB-LS (0)	volta_sgemm_64x64_tn	57767.068455



	Device	Name	Duration
0	Tesla V100-SXM2-32GB-LS (0)	volta_sgemm_128x64_nt	58427.206751
1	Tesla V100-SXM2-32GB-LS (0)	volta_sgemm_32x128_tn	58029.424337
2	Tesla V100-SXM2-32GB-LS (0)	volta_sgemm_128x32_nn	55980.616852
3	Tesla V100-SXM2-32GB-LS (1)	volta_sgemm_32x128_tn	55815.426552
4	Tesla V100-SXM2-32GB-LS (0)	volta_sgemm_64x32_sliced1x4_nt	55182.212475
5	Tesla V100-SXM2-32GB-LS (0)	volta_sgemm_128x32_sliced1x4_nt	54869.676568
6	Tesla V100-SXM2-32GB-LS (1)	volta_sgemm_128x64_nt	54861.393058
7	Tesla V100-SXM2-32GB-LS (1)	volta_sgemm_128x32_nn	53836.994915
8	Tesla V100-SXM2-32GB-LS (0)	volta_sgemm_128x128_nn	52516.189965
9	Tesla V100-SXM2-32GB-LS (1)	volta_sgemm_64x32_sliced1x4_nt	52288.061033

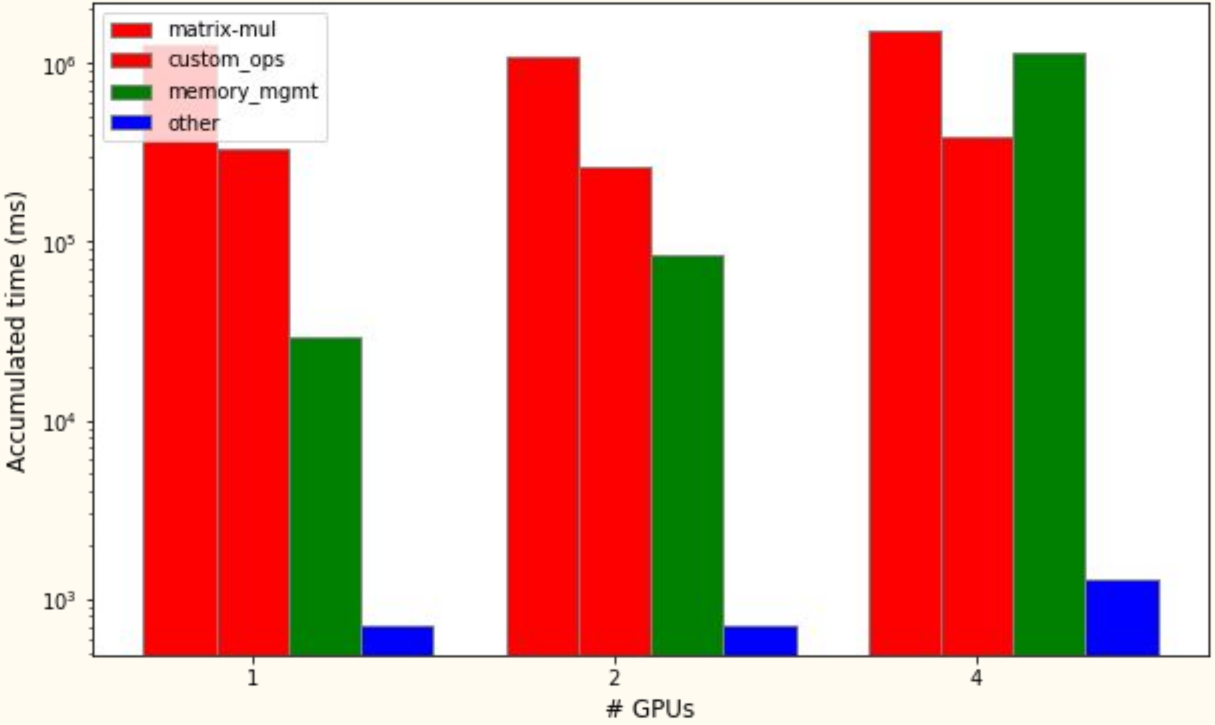
2 GPUs



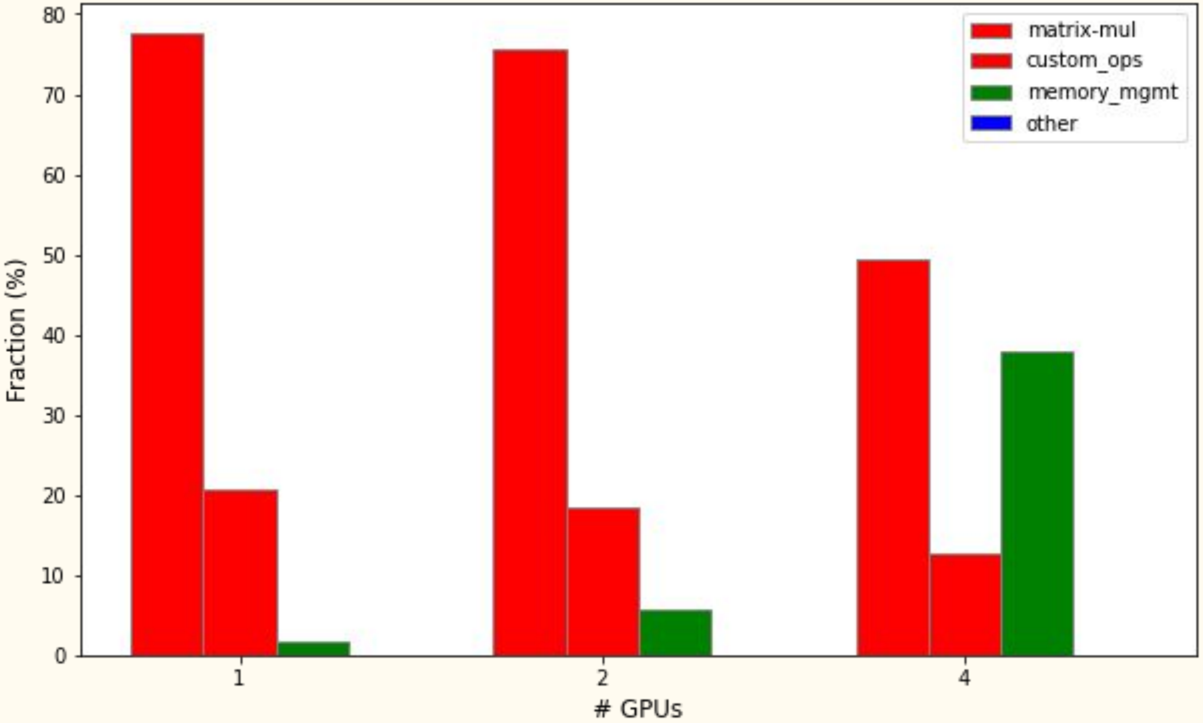
	Device	Name	Duration
0	Tesla V100-SXM2-32GB-LS (0)	ncclReduceRingLLKernel_sum_f32(ncclColl)	157060.957132
1	Tesla V100-SXM2-32GB-LS (2)	ncclReduceRingLLKernel_sum_f32(ncclColl)	156335.110976
2	Tesla V100-SXM2-32GB-LS (3)	ncclReduceRingLLKernel_sum_f32(ncclColl)	155233.676953
3	Tesla V100-SXM2-32GB-LS (1)	ncclReduceRingLLKernel_sum_f32(ncclColl)	152393.073526
4	Tesla V100-SXM2-32GB-LS (1)	volta_sgemm_128x64_tn	127909.728367
5	Tesla V100-SXM2-32GB-LS (2)	ncclBroadcastRingLLKernel_copy_i8(ncclColl)	124997.802063
6	Tesla V100-SXM2-32GB-LS (3)	ncclBroadcastRingLLKernel_copy_i8(ncclColl)	124178.124962
7	Tesla V100-SXM2-32GB-LS (1)	ncclBroadcastRingLLKernel_copy_i8(ncclColl)	123168.228965
8	Tesla V100-SXM2-32GB-LS (3)	volta_sgemm_128x64_tn	122592.641869
9	Tesla V100-SXM2-32GB-LS (0)	ncclBroadcastRingLLKernel_copy_i8(ncclColl)	119301.788980

4 GPUs

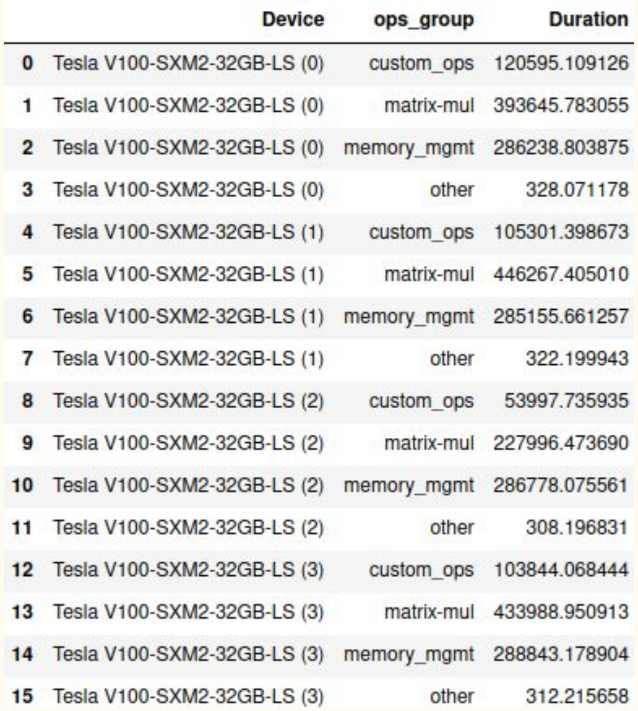
Spending Time in Each Group of Operations: Millisec



Spending Time in Each Group of Operations: Fraction

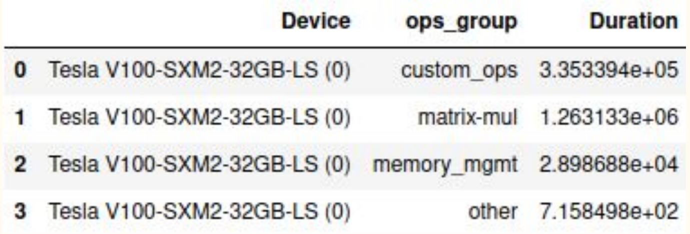


Let’s Dig a little Deeper



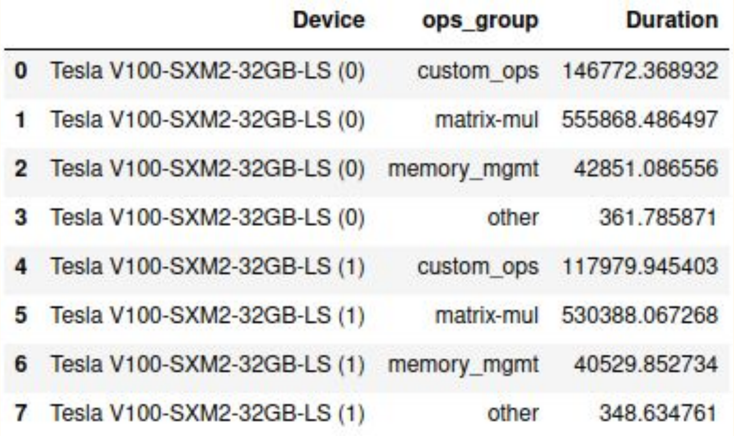
	Device	ops_group	Duration
0	Tesla V100-SXM2-32GB-LS (0)	custom_ops	120595.109126
1	Tesla V100-SXM2-32GB-LS (0)	matrix-mul	393645.783055
2	Tesla V100-SXM2-32GB-LS (0)	memory_mgmt	286238.803875
3	Tesla V100-SXM2-32GB-LS (0)	other	328.071178
4	Tesla V100-SXM2-32GB-LS (1)	custom_ops	105301.398673
5	Tesla V100-SXM2-32GB-LS (1)	matrix-mul	446267.405010
6	Tesla V100-SXM2-32GB-LS (1)	memory_mgmt	285155.661257
7	Tesla V100-SXM2-32GB-LS (1)	other	322.199943
8	Tesla V100-SXM2-32GB-LS (2)	custom_ops	53997.735935
9	Tesla V100-SXM2-32GB-LS (2)	matrix-mul	227996.473690
10	Tesla V100-SXM2-32GB-LS (2)	memory_mgmt	286778.075561
11	Tesla V100-SXM2-32GB-LS (2)	other	308.196831
12	Tesla V100-SXM2-32GB-LS (3)	custom_ops	103844.068444
13	Tesla V100-SXM2-32GB-LS (3)	matrix-mul	433988.950913
14	Tesla V100-SXM2-32GB-LS (3)	memory_mgmt	288843.178904
15	Tesla V100-SXM2-32GB-LS (3)	other	312.215658

4 GPUs



	Device	ops_group	Duration
0	Tesla V100-SXM2-32GB-LS (0)	custom_ops	3.353394e+05
1	Tesla V100-SXM2-32GB-LS (0)	matrix-mul	1.263133e+06
2	Tesla V100-SXM2-32GB-LS (0)	memory_mgmt	2.898688e+04
3	Tesla V100-SXM2-32GB-LS (0)	other	7.158498e+02

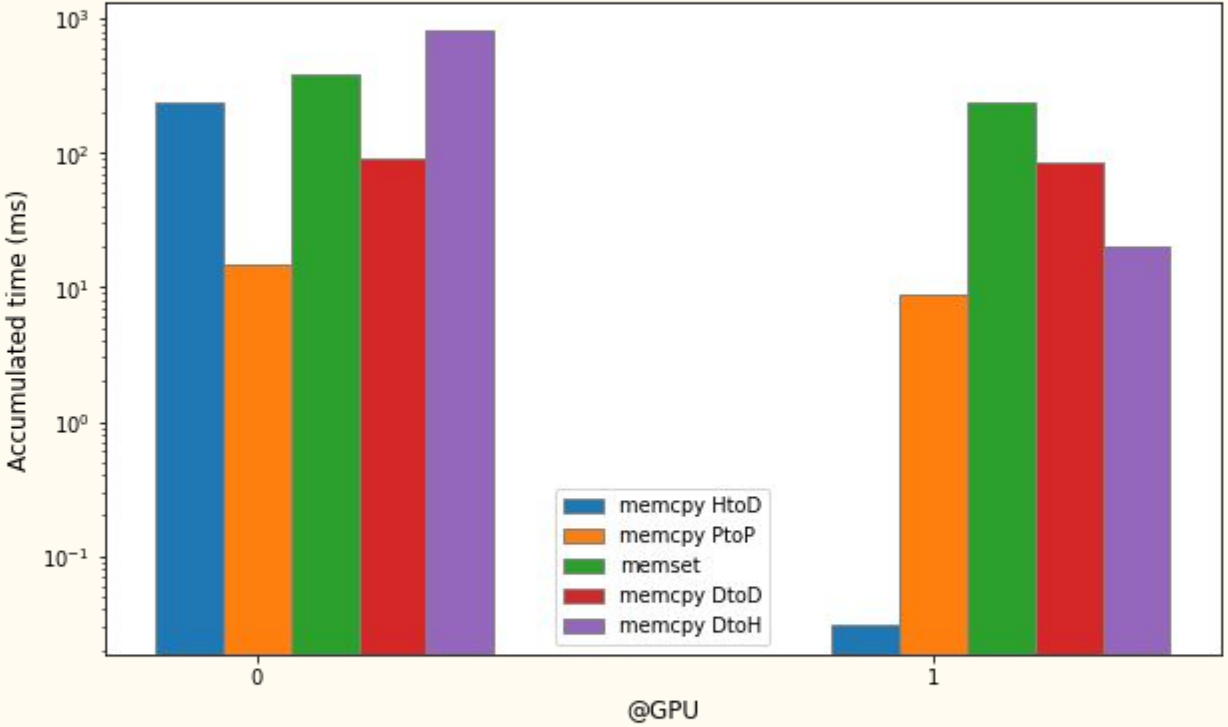
1 GPU



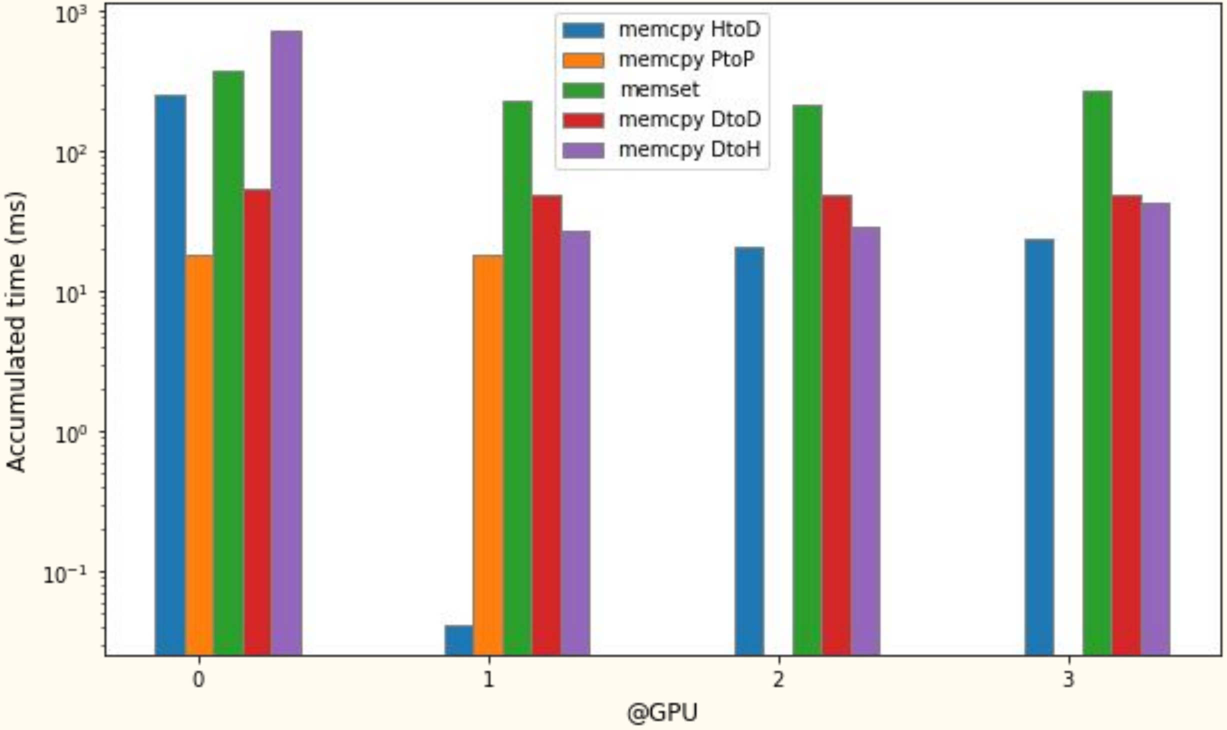
	Device	ops_group	Duration
0	Tesla V100-SXM2-32GB-LS (0)	custom_ops	146772.368932
1	Tesla V100-SXM2-32GB-LS (0)	matrix-mul	555868.486497
2	Tesla V100-SXM2-32GB-LS (0)	memory_mgmt	42851.086556
3	Tesla V100-SXM2-32GB-LS (0)	other	361.785871
4	Tesla V100-SXM2-32GB-LS (1)	custom_ops	117979.945403
5	Tesla V100-SXM2-32GB-LS (1)	matrix-mul	530388.067268
6	Tesla V100-SXM2-32GB-LS (1)	memory_mgmt	40529.852734
7	Tesla V100-SXM2-32GB-LS (1)	other	348.634761

2 GPUs

Communications: 2 GPUs



Communications: 4 GPUs



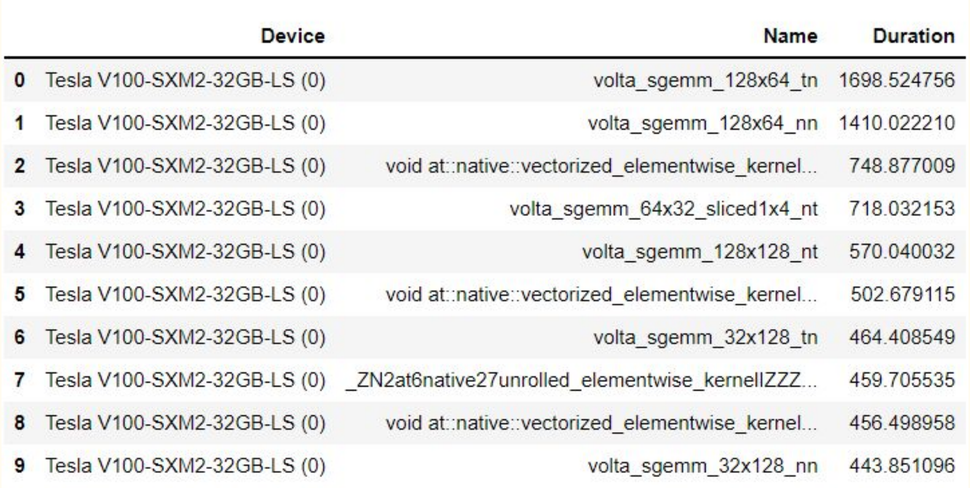
Profiling GPT-2

Configurations

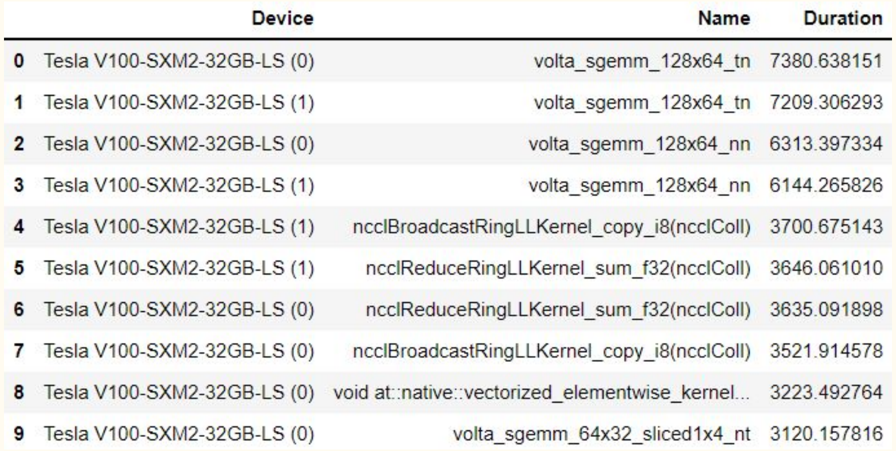
- Batch size: 4
- GPU: 1, 2 and 4
- Model name: DistilGPT2 (the smallest version of GPT2)
- Dataset: yelp_review_full (Text Classification)
- Epoch: 1

GPT-2: Top 10

1 GPU

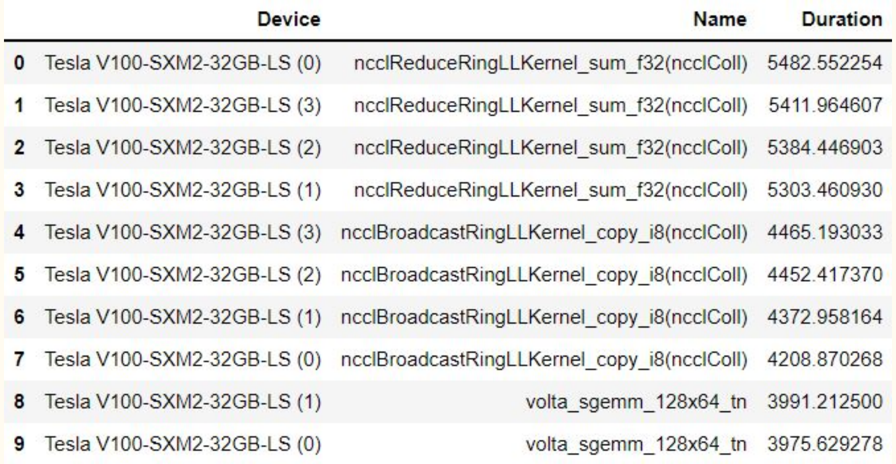


	Device	Name	Duration
0	Tesla V100-SXM2-32GB-LS (0)	volta_sgemm_128x64_tn	1698.524756
1	Tesla V100-SXM2-32GB-LS (0)	volta_sgemm_128x64_nn	1410.022210
2	Tesla V100-SXM2-32GB-LS (0)	void at::native::vectorized_elementwise_kernel...	748.877009
3	Tesla V100-SXM2-32GB-LS (0)	volta_sgemm_64x32_sliced1x4_nt	718.032153
4	Tesla V100-SXM2-32GB-LS (0)	volta_sgemm_128x128_nt	570.040032
5	Tesla V100-SXM2-32GB-LS (0)	void at::native::vectorized_elementwise_kernel...	502.679115
6	Tesla V100-SXM2-32GB-LS (0)	volta_sgemm_32x128_tn	464.408549
7	Tesla V100-SXM2-32GB-LS (0)	_ZN2at6native27unrolled_elementwise_kernelIZZZ...	459.705535
8	Tesla V100-SXM2-32GB-LS (0)	void at::native::vectorized_elementwise_kernel...	456.498958
9	Tesla V100-SXM2-32GB-LS (0)	volta_sgemm_32x128_nn	443.851096



	Device	Name	Duration
0	Tesla V100-SXM2-32GB-LS (0)	volta_sgemm_128x64_tn	7380.638151
1	Tesla V100-SXM2-32GB-LS (1)	volta_sgemm_128x64_tn	7209.306293
2	Tesla V100-SXM2-32GB-LS (0)	volta_sgemm_128x64_nn	6313.397334
3	Tesla V100-SXM2-32GB-LS (1)	volta_sgemm_128x64_nn	6144.265826
4	Tesla V100-SXM2-32GB-LS (1)	ncclBroadcastRingLLKernel_copy_i8(ncclColl)	3700.675143
5	Tesla V100-SXM2-32GB-LS (1)	ncclReduceRingLLKernel_sum_f32(ncclColl)	3646.061010
6	Tesla V100-SXM2-32GB-LS (0)	ncclReduceRingLLKernel_sum_f32(ncclColl)	3635.091898
7	Tesla V100-SXM2-32GB-LS (0)	ncclBroadcastRingLLKernel_copy_i8(ncclColl)	3521.914578
8	Tesla V100-SXM2-32GB-LS (0)	void at::native::vectorized_elementwise_kernel...	3223.492764
9	Tesla V100-SXM2-32GB-LS (0)	volta_sgemm_64x32_sliced1x4_nt	3120.157816

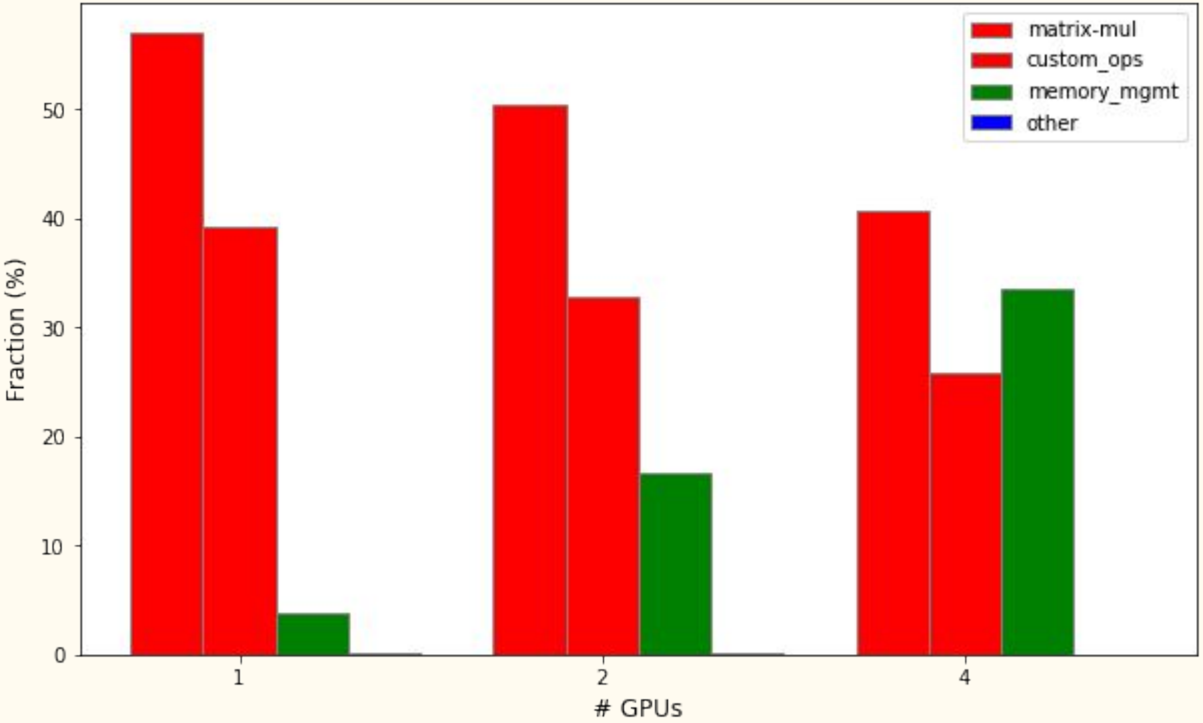
2 GPUs



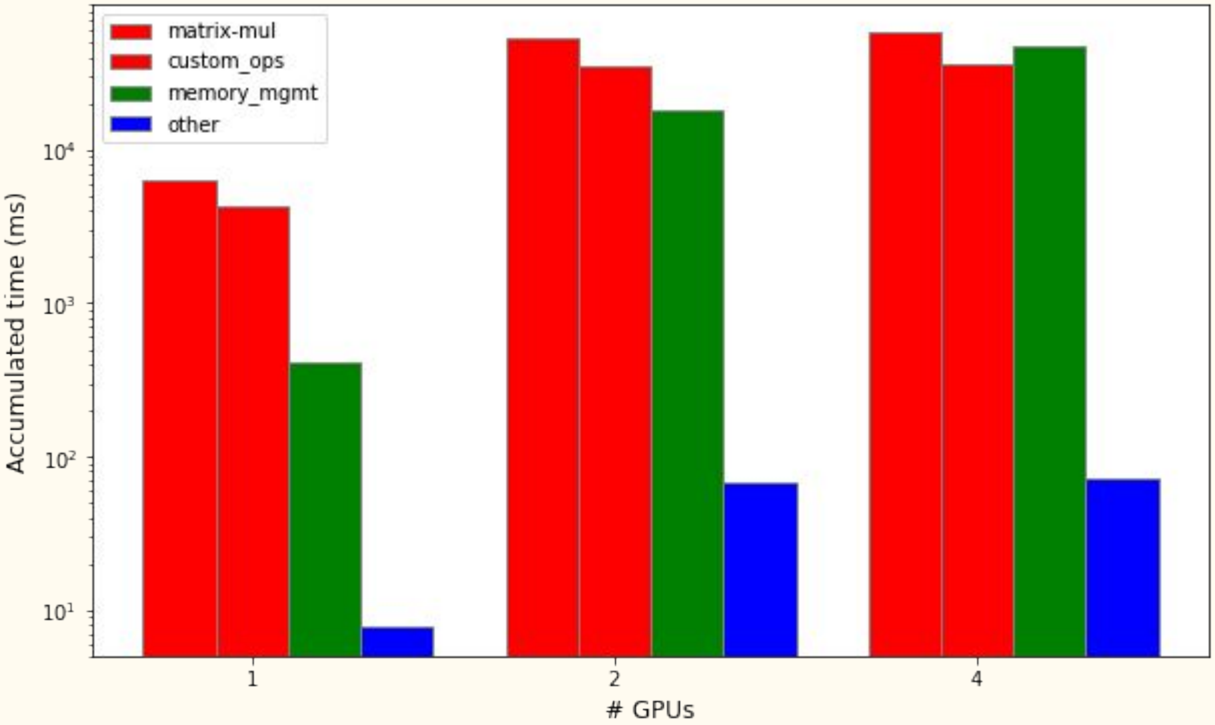
	Device	Name	Duration
0	Tesla V100-SXM2-32GB-LS (0)	ncclReduceRingLLKernel_sum_f32(ncclColl)	5482.552254
1	Tesla V100-SXM2-32GB-LS (3)	ncclReduceRingLLKernel_sum_f32(ncclColl)	5411.964607
2	Tesla V100-SXM2-32GB-LS (2)	ncclReduceRingLLKernel_sum_f32(ncclColl)	5384.446903
3	Tesla V100-SXM2-32GB-LS (1)	ncclReduceRingLLKernel_sum_f32(ncclColl)	5303.460930
4	Tesla V100-SXM2-32GB-LS (3)	ncclBroadcastRingLLKernel_copy_i8(ncclColl)	4465.193033
5	Tesla V100-SXM2-32GB-LS (2)	ncclBroadcastRingLLKernel_copy_i8(ncclColl)	4452.417370
6	Tesla V100-SXM2-32GB-LS (1)	ncclBroadcastRingLLKernel_copy_i8(ncclColl)	4372.958164
7	Tesla V100-SXM2-32GB-LS (0)	ncclBroadcastRingLLKernel_copy_i8(ncclColl)	4208.870268
8	Tesla V100-SXM2-32GB-LS (1)	volta_sgemm_128x64_tn	3991.212500
9	Tesla V100-SXM2-32GB-LS (0)	volta_sgemm_128x64_tn	3975.629278

4 GPUs

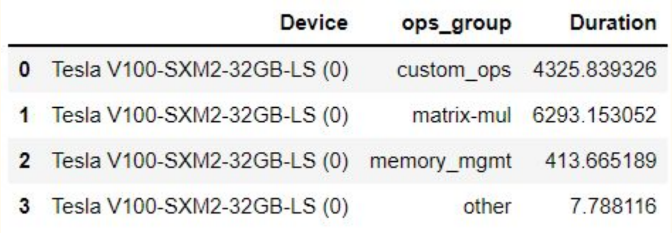
Spending Time in Each Group of Operations: Fraction



Spending Time in Each Group of Operations: Millisec

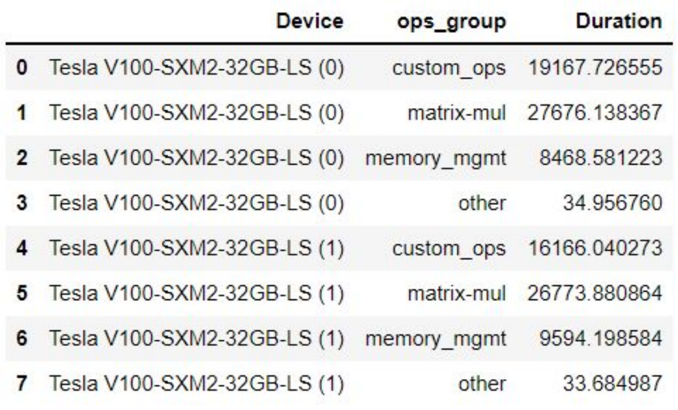


Let’s Dig a little Deeper



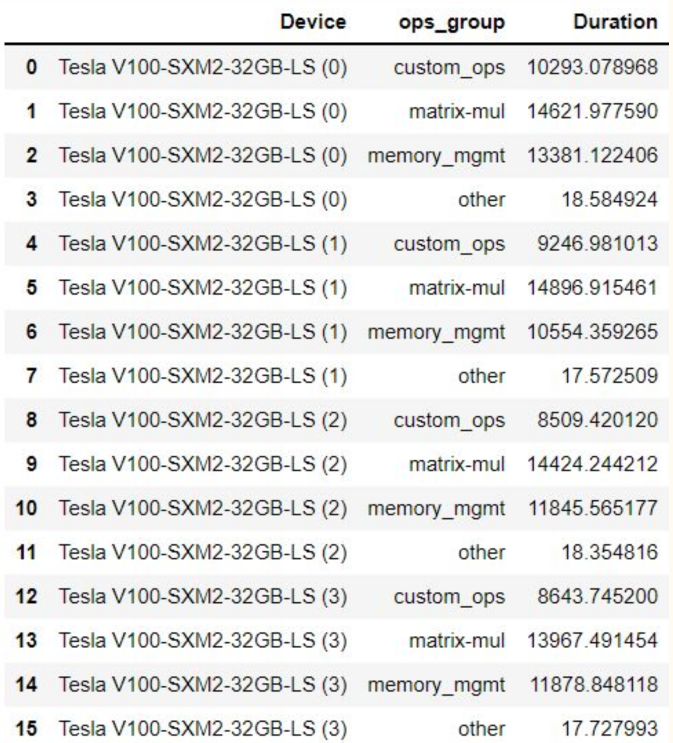
	Device	ops_group	Duration
0	Tesla V100-SXM2-32GB-LS (0)	custom_ops	4325.839326
1	Tesla V100-SXM2-32GB-LS (0)	matrix-mul	6293.153052
2	Tesla V100-SXM2-32GB-LS (0)	memory_mgmt	413.665189
3	Tesla V100-SXM2-32GB-LS (0)	other	7.788116

1 GPU



	Device	ops_group	Duration
0	Tesla V100-SXM2-32GB-LS (0)	custom_ops	19167.726555
1	Tesla V100-SXM2-32GB-LS (0)	matrix-mul	27676.138367
2	Tesla V100-SXM2-32GB-LS (0)	memory_mgmt	8468.581223
3	Tesla V100-SXM2-32GB-LS (0)	other	34.956760
4	Tesla V100-SXM2-32GB-LS (1)	custom_ops	16166.040273
5	Tesla V100-SXM2-32GB-LS (1)	matrix-mul	26773.880864
6	Tesla V100-SXM2-32GB-LS (1)	memory_mgmt	9594.198584
7	Tesla V100-SXM2-32GB-LS (1)	other	33.684987

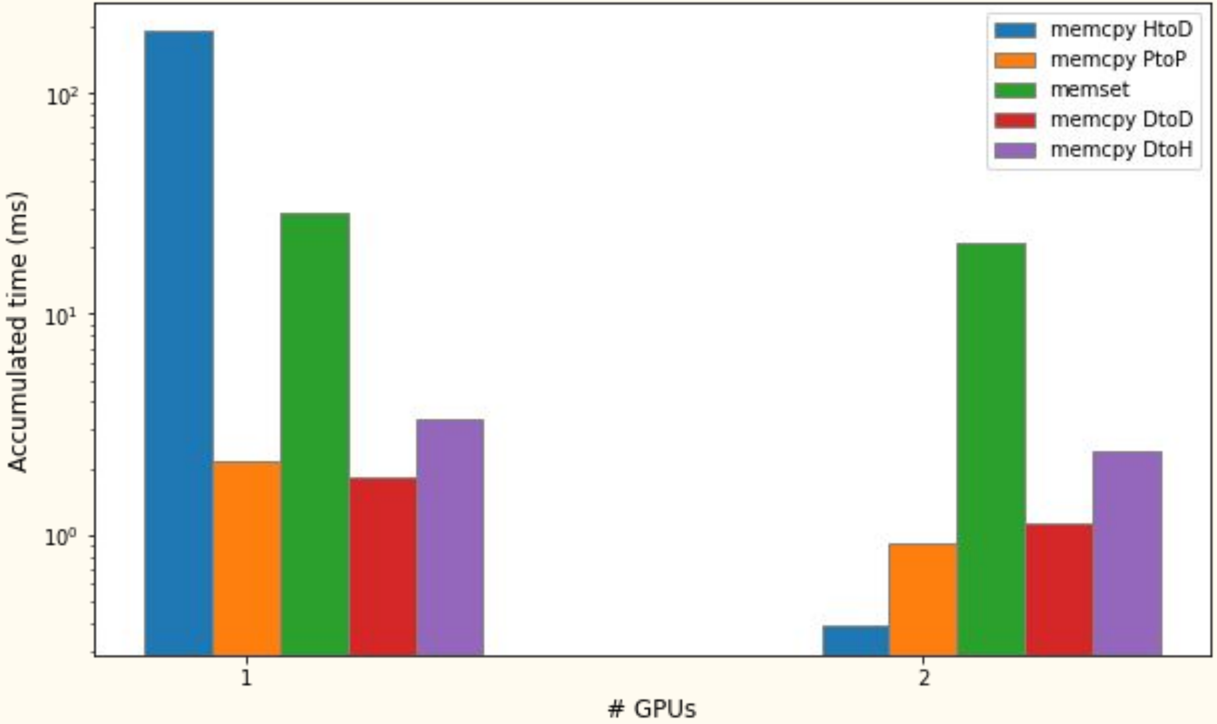
2 GPUs



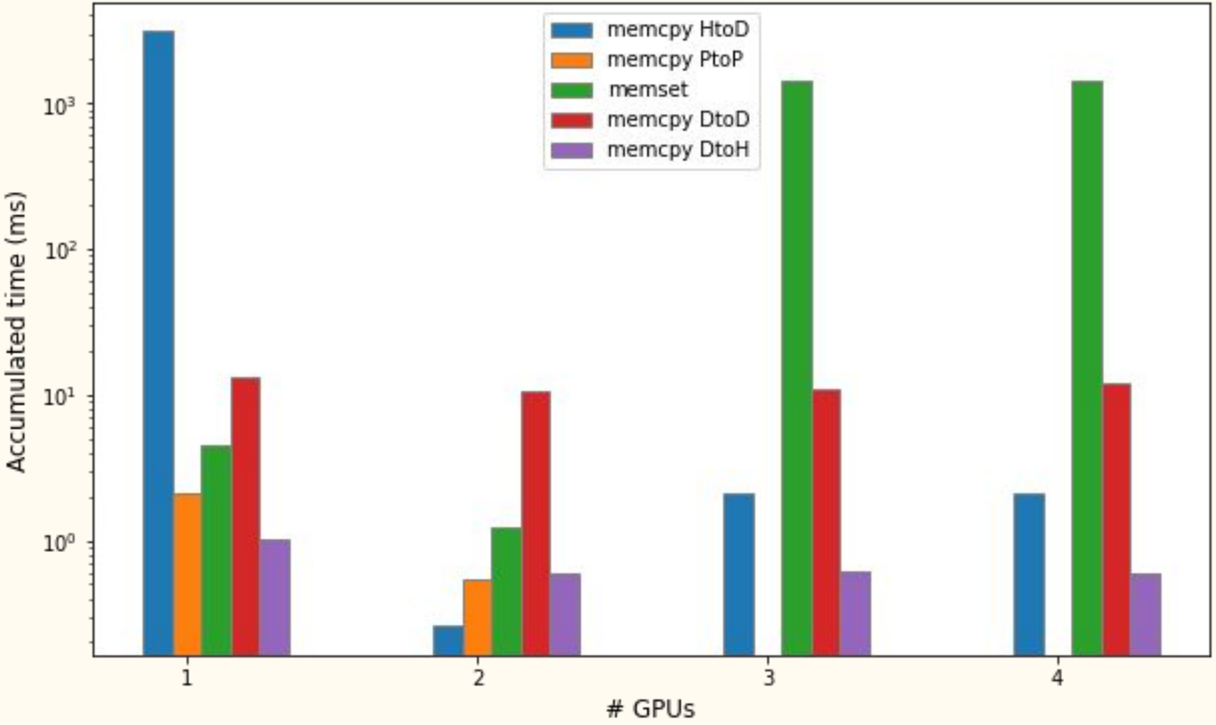
	Device	ops_group	Duration
0	Tesla V100-SXM2-32GB-LS (0)	custom_ops	10293.078968
1	Tesla V100-SXM2-32GB-LS (0)	matrix-mul	14621.977590
2	Tesla V100-SXM2-32GB-LS (0)	memory_mgmt	13381.122406
3	Tesla V100-SXM2-32GB-LS (0)	other	18.584924
4	Tesla V100-SXM2-32GB-LS (1)	custom_ops	9246.981013
5	Tesla V100-SXM2-32GB-LS (1)	matrix-mul	14896.915461
6	Tesla V100-SXM2-32GB-LS (1)	memory_mgmt	10554.359265
7	Tesla V100-SXM2-32GB-LS (1)	other	17.572509
8	Tesla V100-SXM2-32GB-LS (2)	custom_ops	8509.420120
9	Tesla V100-SXM2-32GB-LS (2)	matrix-mul	14424.244212
10	Tesla V100-SXM2-32GB-LS (2)	memory_mgmt	11845.565177
11	Tesla V100-SXM2-32GB-LS (2)	other	18.354816
12	Tesla V100-SXM2-32GB-LS (3)	custom_ops	8643.745200
13	Tesla V100-SXM2-32GB-LS (3)	matrix-mul	13967.491454
14	Tesla V100-SXM2-32GB-LS (3)	memory_mgmt	11878.848118
15	Tesla V100-SXM2-32GB-LS (3)	other	17.727993

4 GPUs

Communications: 2 GPUs



Communications: 4 GPUs



spec

Hardware

- GPU: Tesla V100 x 8
- CPU: Intel(R) Xeon(R) CPU E5-2698 v4 @ 2.20GHz
- RAM: 503 GB

Software

- nvprof 10.2.89
- python 3.6.
- CUDA NVIDIA-SMI 450.119.04
- Driver Version: 450.119.04
- CUDA Version: 11.0
- Running on (Docker) Container

Conclusions

- The horizontal scaling for GPU training is highly non-linear
- Latency came from communication overhead and increasing as # of GPUs grow
- Found an asymmetry in spending time of memory management

QA