GPUMemNet: Deep Learning-based Estimation of GPU Memory Requirement for Neural Network Training Tasks

Ehsan Yousefzadeh-Asl-Miandoab[1] (ehyo@itu.dk), Reza Karimzadeh[2], Bulat Ibragimov[2], Pınar Tözün[1]
(1) IT University of Copenhagen, (2) University of Copenhagen

1

GPU Underutilization: Causes and Opportunities

Real-world clusters exhibit only ~50% GPU utilization *

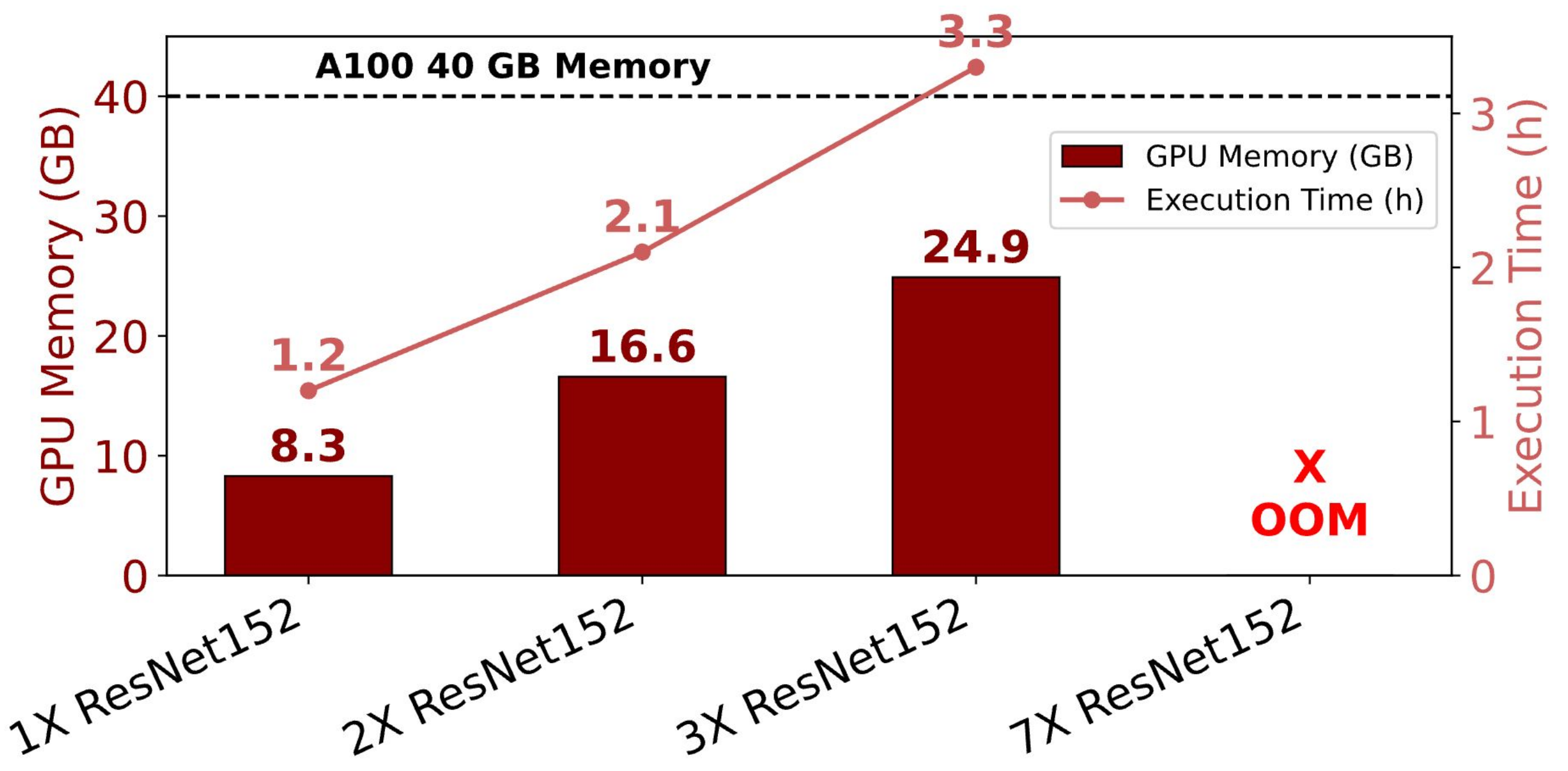
1- GPUs' lack of **fine-grain sharing** and **virtual memory**
2- **Exclusive** GPU assignment by resource managers
3- **Black box** view of tasks and GPUs

Collocating tasks together increase GPU utilization!

* Yanjie Gao et al. "**An Empirical Study on Low GPU Utilization of Deep Learning Jobs**," ICSE'24.

2

OOM Crashes & Interference!



GPU memory estimation is essential before robust collocating.

3

Estimating GPU memory — easier said than done!

- GPU memory optimizations = complex data patterns
  → analytics can't keep up.

Machine Learning excels at pattern recognition.

4

ML-based Estimator Challenges

1- Building a representative dataset that generalizes well!

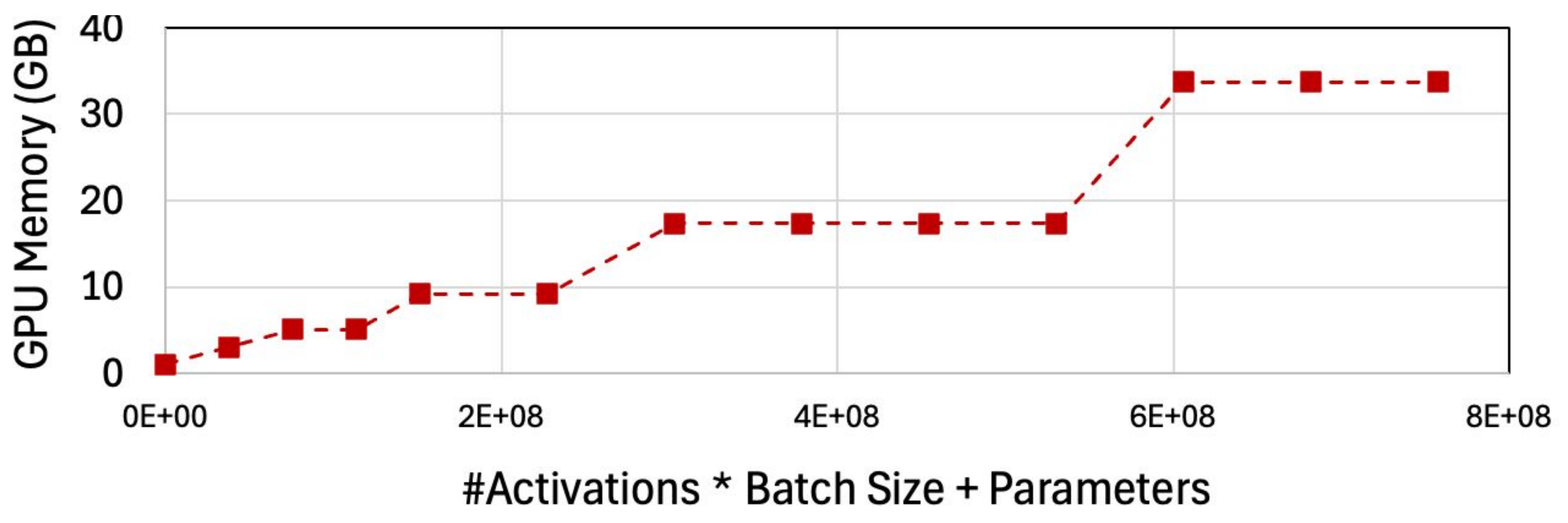
2- Formulating the problem effectively!

5

Dataset Building: What Matters

1- Focus on architecture not the model types.
2- Representative range of features.
3- Uniform feature distribution.
4- Diversity of shapes within an architecture.
5- Diversity of layers in practical architectures.
6- Varying input and output sizes.

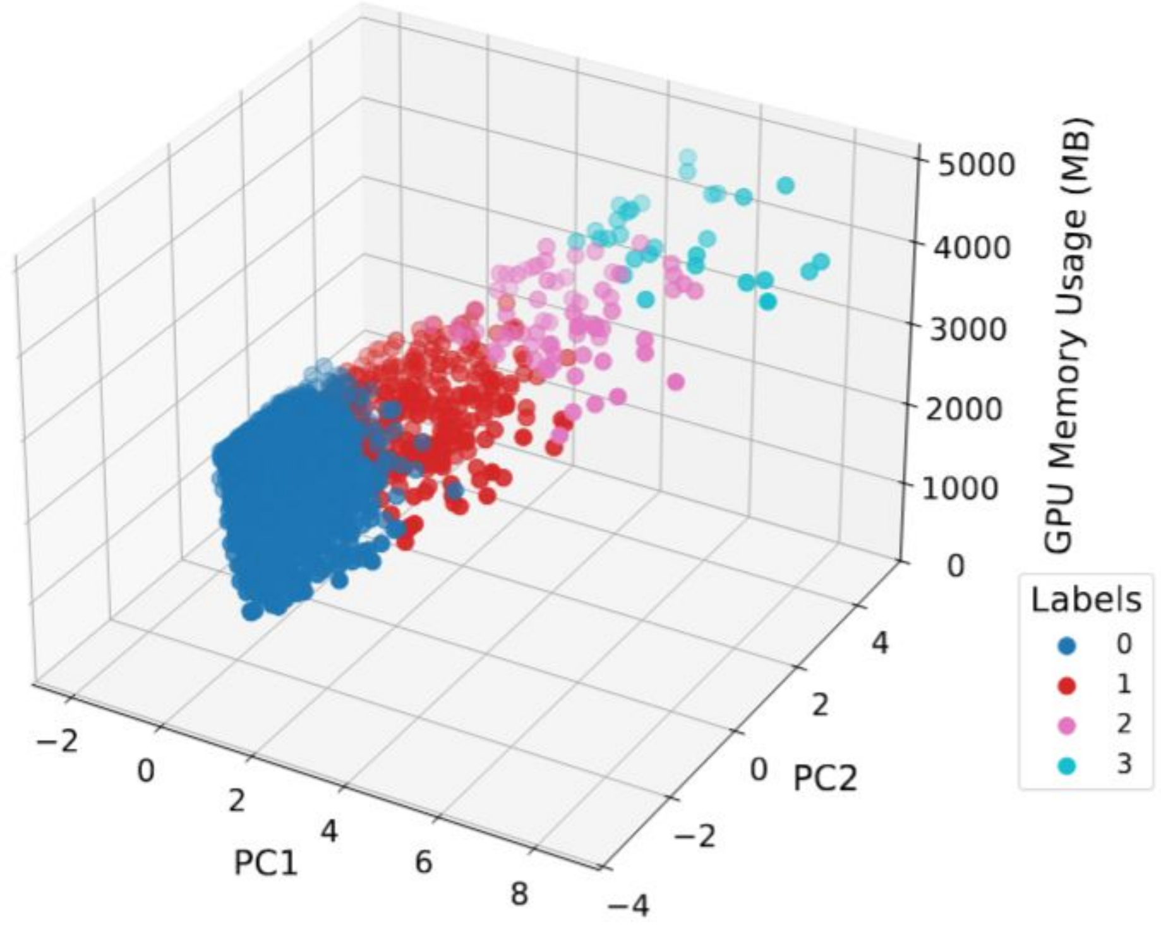
Staircase growth → a classification problem



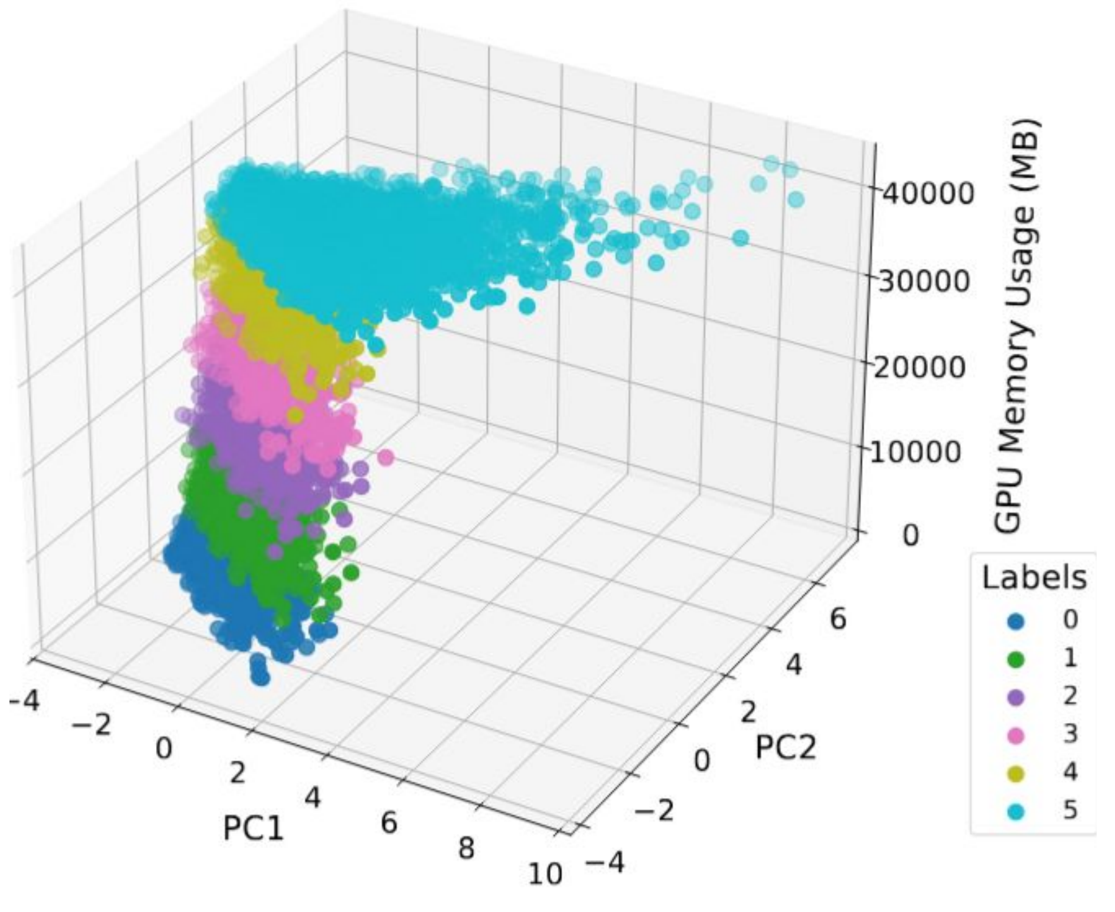
6

PCA Analysis

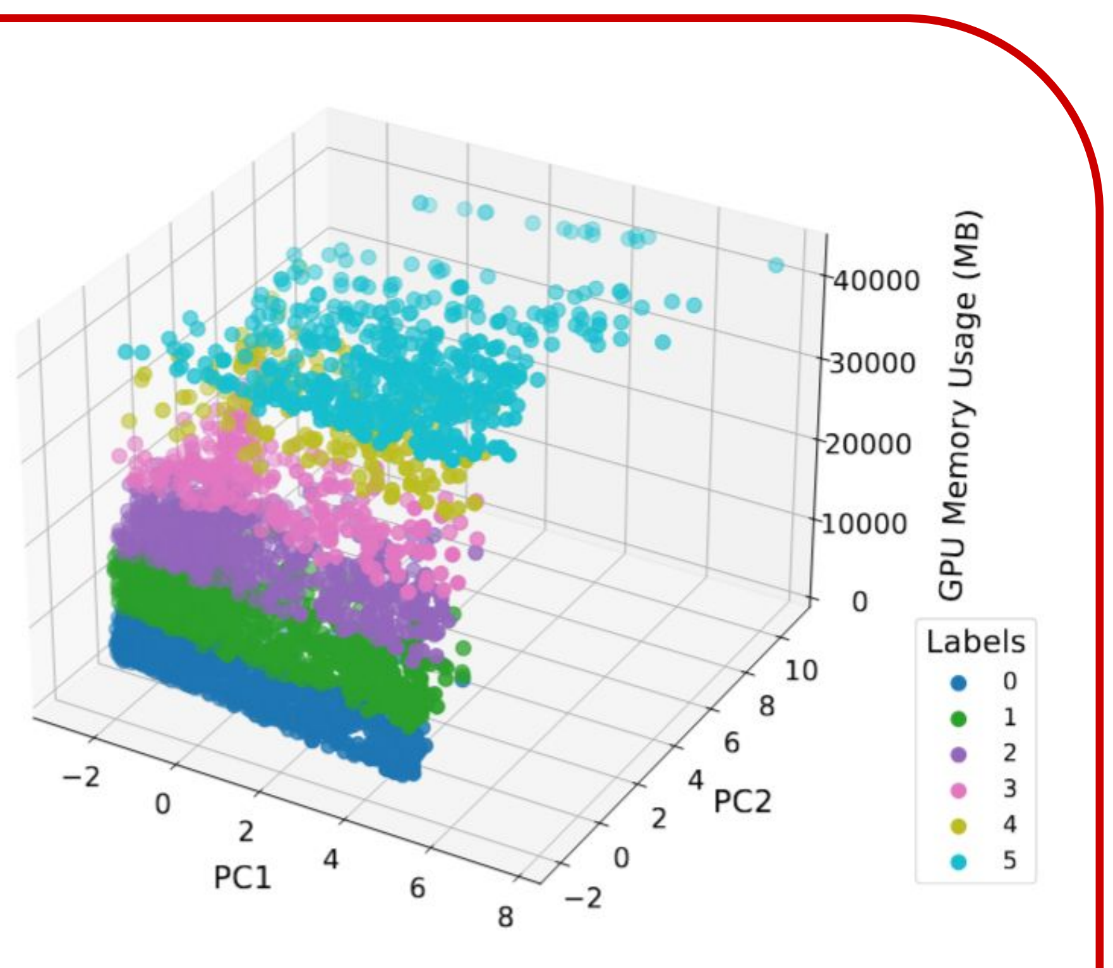
Curated dataset:
- 3K MLPs
- 9K CNNs
- 5K Transformers



MLP



CNN

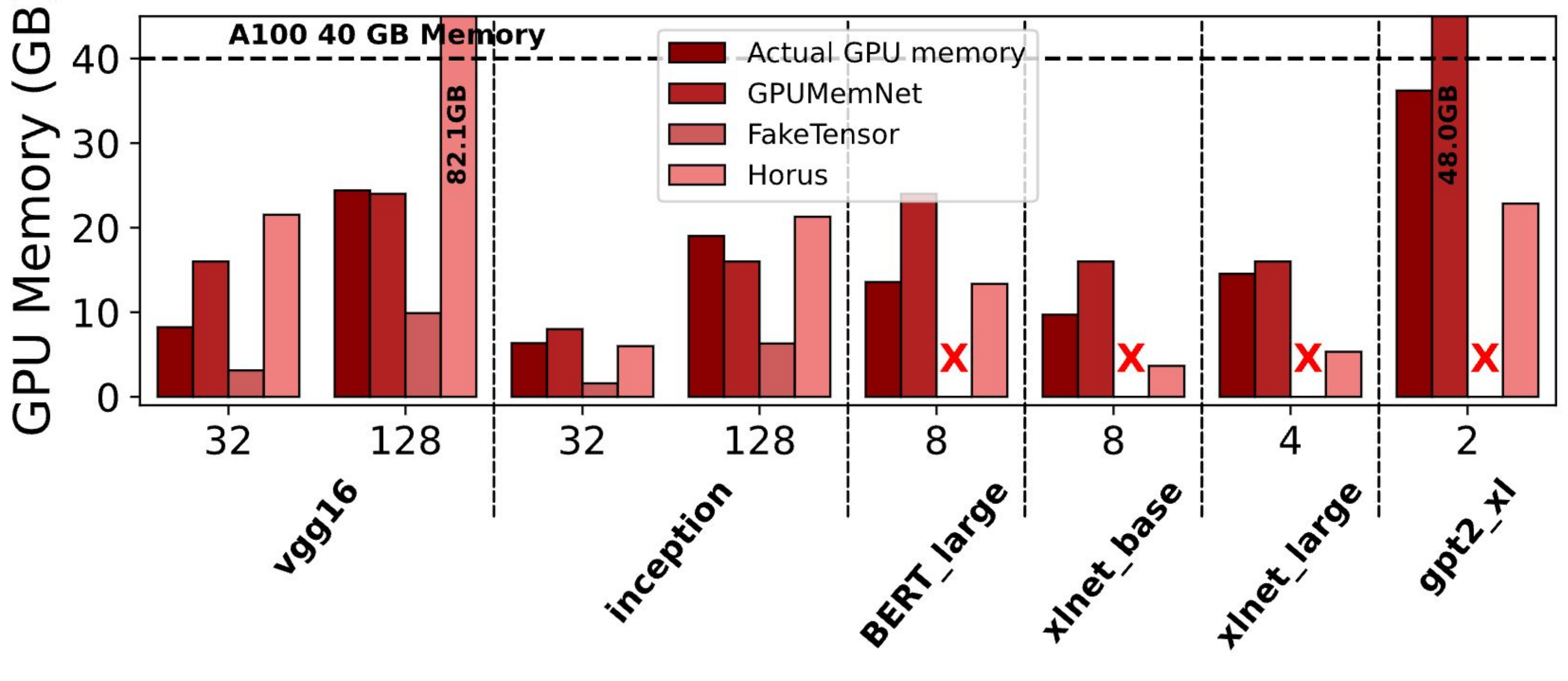


Transformer

7

Trained MLPs & Evaluation

Dataset	Range	Accuracy	F1-Score
MLPs	2GB	0.97	0.96
	1GB	0.95	0.93
CNNs	8GB	0.83	0.83
Transformers	8GB	0.88	0.88



Batch Size, Workload (from top to bottom)

See the code on GitHub!

IT UNIVERSITY OF COPENHAGEN

itu.dk

rad.itu.dk