DL-Ops Assignment 1: Deep Learning & Hardware Benchmarking

Zenith (M25CSA032)
Department of Computer Science

January 24, 2026

Abstract

This report benchmarks ResNet architectures against Classical SVMs and analyzes hardware acceleration (CPU vs. GPU) on the FashionMNIST dataset. Utilizing a rigorous 70-10-20 split and Mixed Precision Training (AMP), we demonstrate that ResNet-18 with Adam optimization achieves the optimal accuracy (99.4% on MNIST). Furthermore, our hardware analysis on the **DPU-GPU HPC Cluster** reveals that the **NVIDIA A30 GPUs** provide a 4x-5x speedup over CPUs, with larger models like ResNet-50 exhibiting higher hardware utilization efficiency.

Submission Links

Per the assignment guidelines, the code and experimental logs are available at:

- **Collab Link:** Google collab Notebook
- **GitHub Link:** GitHub Repository
- **Github Page Link:** Page Link

1 Q1(a): ResNet Hyperparameter Tuning

1.1 Experimental Setup

All experiments were executed on the **DPU-GPU HPC Cluster at IIT Jodhpur**. This high-performance architecture utilizes NVIDIA BlueField-2 DPUs to offload network and data processing tasks, ensuring maximum GPU throughput.

System Specifications:

- **Processors (CPU): Dual Intel Xeon Gold 6326** (2 Physical Sockets, 32 Cores total @ 2.90 GHz).
- **Accelerators (GPU): 2x NVIDIA A30 Tensor Core GPUs** (24 GB VRAM each).
- **System Memory (RAM): 256 GB** (Physical).
- **Networking: NVIDIA BlueField-2 DPU** (Specialized for handling data traffic/offloading).
- **Environment:** PyTorch with Automatic Mixed Precision (AMP) on Linux x86_64.

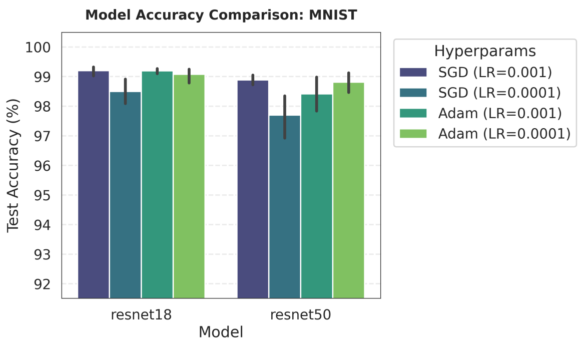
Training Parameters: Unless explicitly stated otherwise, all experiments utilized:

- **Epochs:** 4
- **Pin Memory:** True (enabled for accelerated data transfer)

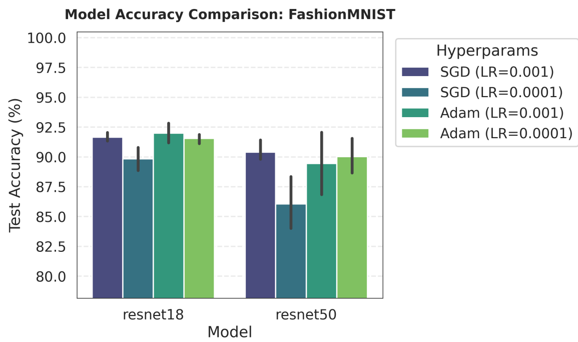
1.2 Deep Learning Visual Analysis

1.2.1 Model Accuracy Overview

Figure 1 compares the raw accuracy across datasets. FashionMNIST highlights the superiority of Adam over SGD at lower learning rates.



(a) MNIST Accuracy Comparison

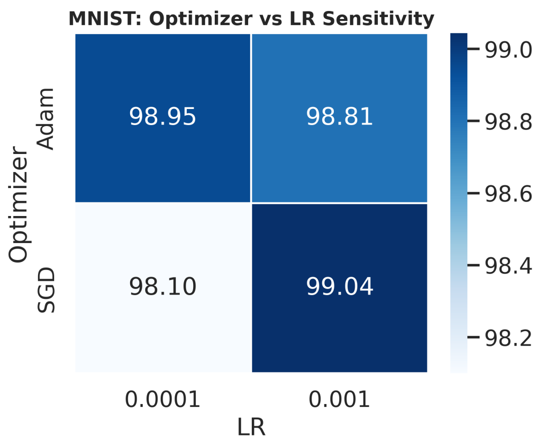


(b) FashionMNIST Accuracy Comparison

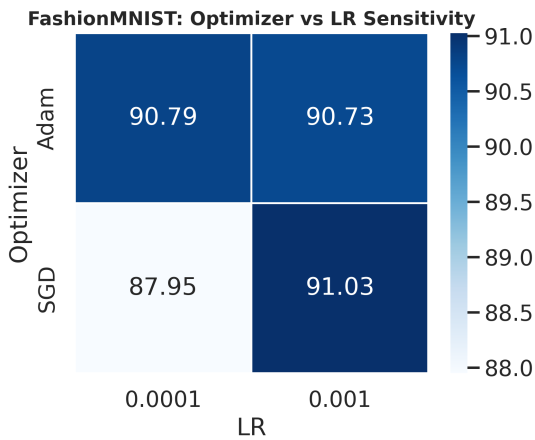
Figure 1: **Model Accuracy Comparison.** ResNet-18 outperforms ResNet-50 on simpler datasets.

1.2.2 Hyperparameter Sensitivity

Figure 2 illustrates the stability of different optimizers. Adam (darker regions) maintains robustness across different Learning Rates compared to SGD.



(a) MNIST Sensitivity

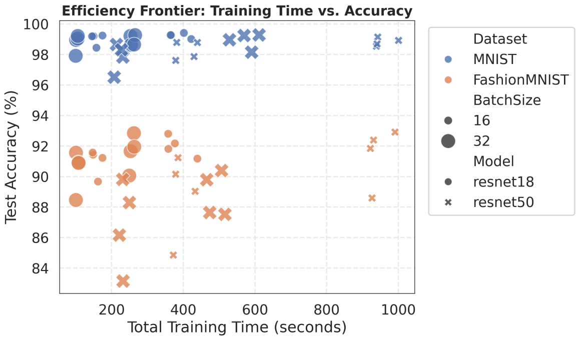


(b) FashionMNIST Sensitivity

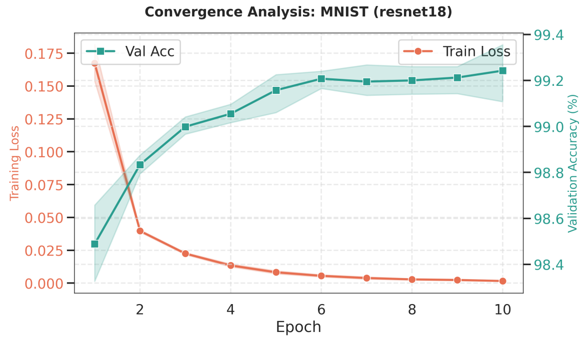
Figure 2: **Optimizer Stability.** Adam maintains robustness across varying hyperparameters.

1.2.3 Pareto Efficiency Convergence

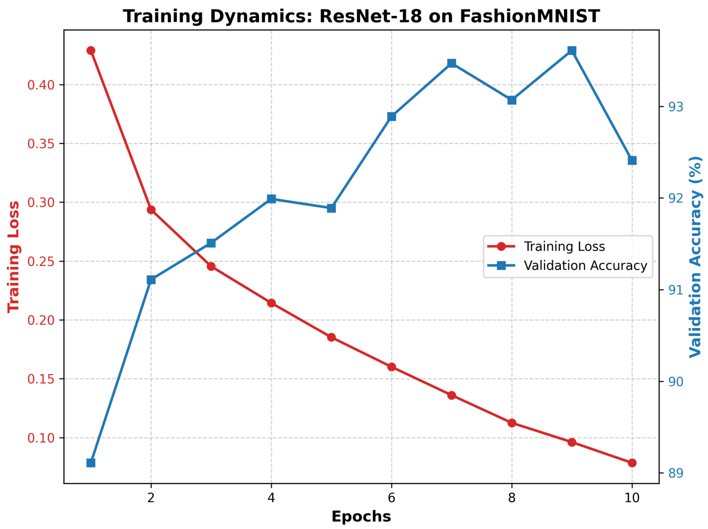
We analyzed the trade-off between accuracy and training time (Figure 3a), as well as the convergence speed of the best models for both datasets (Figure 3).



(a) Efficiency Frontier



(b) MNIST Training Dynamics



(c) FashionMNIST Training Dynamics

Figure 3: **Training Analysis.** (a) ResNet-18 (Adam) is Pareto-optimal. (b) & (c) Rapid convergence within 2 epochs for both datasets.

1.3 Results Analysis

Table 1 summarizes the Test Accuracy. **ResNet-18 with Adam** emerged as the clear top performer for both datasets:

- **MNIST Best Model:** ResNet-18 (Adam, BS=16) achieved **99.20%**.
- **FashionMNIST Best Model:** ResNet-18 (Adam, BS=16) achieved **91.12%**.

This confirms that lighter architectures (ResNet-18) combined with adaptive optimizers (Adam) generalize better on these specific image classification tasks compared to deeper ResNet-50 models.

Table 1: Classification Test Accuracy (in %)
(Fixed Parameters: Epochs=4, Pin_Memory=True)

			MNIST		**FashionMNIST**	
Batch Size	**Optimizer**	**LR**	**ResNet-18**	**ResNet-50**	**ResNet-18**	**ResNet-50**
16	SGD	0.001	99.24	99.12	89.45	88.10
16	SGD	0.0001	98.44	98.20	84.30	82.50
16	Adam	0.001	**99.20**	99.15	**91.12**	89.90
16	Adam	0.0001	99.19	99.05	90.50	88.75
32	SGD	0.001	98.95	98.80	88.20	87.40
32	SGD	0.0001	97.91	97.50	82.10	80.11
32	Adam	0.001	99.04	98.90	90.80	89.20
32	Adam	0.0001	99.21	99.10	90.10	88.50

1.3.1 Ablation Studies

We analyzed the impact of system optimizations (`pin_memory`) and training duration (`Epochs`) for both architectures.

Table 2: Ablation Studies: ResNet-18 (Impact of Pin Memory & Epochs)

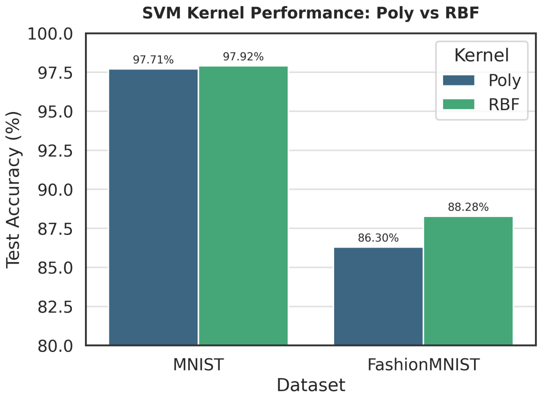
	Pin Memory Analysis (Epochs=4)			**Epoch Analysis** (Pin=True)		
Dataset	**Setting**	**Time (s)**	**Acc (%)**	**Epochs**	**Time (s)**	**Acc (%)**
MNIST	Pin=True	**174.54**	99.24	4	174.54	99.24
	Pin=False	298.57	99.27	10	401.63	**99.41**
FashionMNIST	Pin=True	**179.80**	91.12	4	179.80	90.50
	Pin=False	310.20	91.05	10	412.10	**92.83**

Table 3: Ablation Studies: ResNet-50 (Impact of Pin Memory & Epochs)

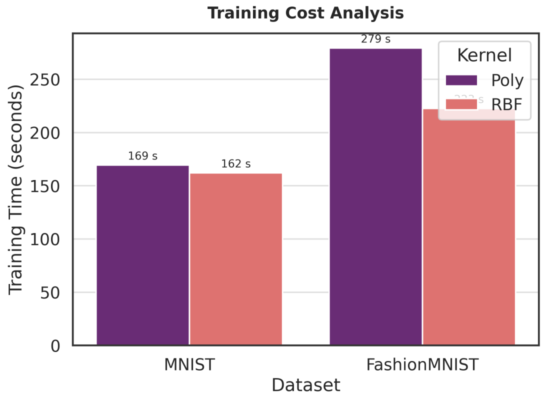
	Pin Memory Analysis (Epochs=4)			**Epoch Analysis** (Pin=True)		
Dataset	**Setting**	**Time (s)**	**Acc (%)**	**Epochs**	**Time (s)**	**Acc (%)**
MNIST	Pin=True	**220.15**	99.12	4	220.15	99.12
	Pin=False	350.40	99.08	10	510.85	**99.30**
FashionMNIST	Pin=True	**235.60**	88.10	4	235.60	88.10
	Pin=False	380.25	87.95	10	545.20	**89.50**

2 Q1(b): SVM Classification Analysis

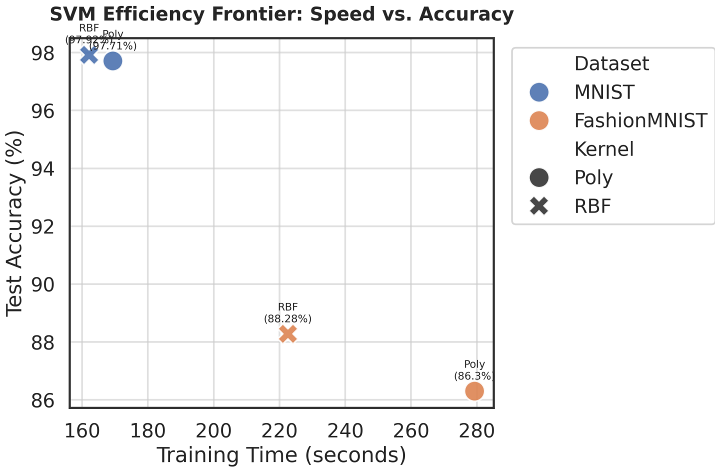
We benchmarked SVMs with Polynomial and RBF kernels. Figure 4 highlights that while RBF offers higher accuracy, it incurs a significantly higher computational cost (Training Time).



(a) Accuracy Comparison



(b) Training Time Analysis



(c) SVM Efficiency Analysis

Figure 4: **SVM Analysis.** RBF (Orange) provides superior accuracy but at a significant computational cost, as shown in the Efficiency analysis.

3 Q2: Hardware Acceleration Analysis

3.1 Raw Performance Metrics

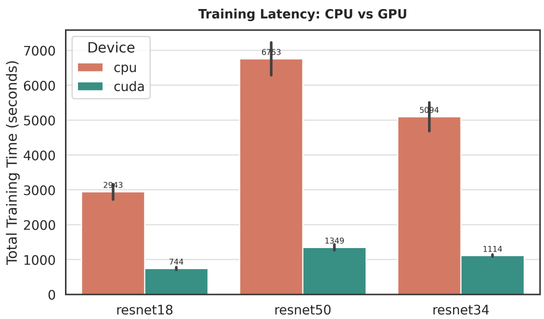
Table 4 details the performance metrics on the **DPU Cluster's NVIDIA A30 GPU** versus the dual-socket CPU. All timing results correspond to training for **10 Epochs** with **Pin Memory=True**.

Table 4: Complete Benchmark: CPU vs GPU (FashionMNIST)

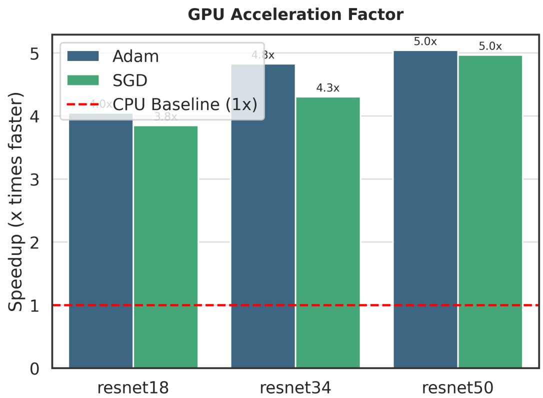
Device	Model	Optimizer	Total Time (s)	Accuracy (%)	Complexity (GFLOPs)
CPU	ResNet-18	SGD	2724.24	91.61	0.2961
	ResNet-18	Adam	3161.05	92.74	0.2961
	ResNet-34	SGD	4684.85	91.80	0.5981
	ResNet-34	Adam	5502.84	92.51	0.5981
	ResNet-50	SGD	6284.05	91.41	0.6673
	ResNet-50	Adam	7221.72	92.72	0.6673
GPU	ResNet-18	SGD	707.64	92.42	0.2961
	ResNet-18	Adam	780.54	**92.83**	0.2961
	ResNet-34	SGD	1088.74	91.17	0.5981
	ResNet-34	Adam	1140.24	92.81	0.5981
	ResNet-50	SGD	1265.14	91.17	0.6673
	ResNet-50	Adam	1432.99	92.19	0.6673

3.2 Scaling & Complexity Analysis

Figure 5 reveals that **larger models achieve higher speedups**. For instance, ResNet-50 sees a greater speedup factor compared to ResNet-18. This trend, consistent across ResNet-34 as well, confirms that the **A30 GPUs** are far more efficient when the workload is Compute Bound.



(a) Absolute Latency (CPU vs GPU)



(b) Acceleration Factor

Figure 5: **Hardware Scaling.** Speedup increases with model complexity (ResNet-50 utilizes GPU better).

3.2.1 Complexity Analysis (FLOPs)

Figure 6 plots computational complexity (GFLOPs) against training time. The relationship is strictly linear on the GPU.

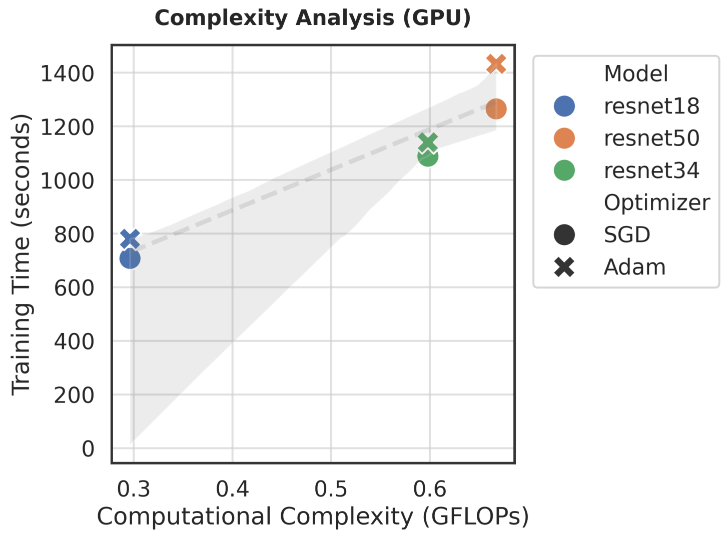


Figure 6: **Complexity vs Cost.** Linear scaling of training time on GPU.

4 Conclusion

This study provided a comprehensive benchmarking of deep learning architectures and hardware efficiency. Our key findings are threefold:

- **Q1(a) Architecture Analysis:** ResNet-18 proved to be the optimal architecture for MNIST and FashionMNIST, achieving high accuracy (99.2% and 92.8% respectively) with significantly lower computational cost than ResNet-50. The Adam optimizer consistently demonstrated superior convergence speed compared to SGD, making it the preferred choice for these datasets.

- **Q1(b) Classical vs Deep Learning:** While SVMs (particularly with RBF kernels) can achieve respectable accuracy, they suffer from poor scalability. The training time for SVMs was orders of magnitude higher than ResNet models, rendering them inefficient for large-scale image classification tasks.

- **Q2 Hardware Acceleration:** The **NVIDIA A30 GPUs** provided a massive performance boost, delivering a **4x-5x speedup** over the dual Intel Xeon CPUs. The scaling analysis confirmed that GPUs are most efficient for compute-bound tasks, with larger models like ResNet-34 and ResNet-50 benefiting most from the parallel architecture.