Performance Comparison of Reversible Vision Transformer Models

Models of ViT-small, Reversible ViT-small and ViT-small with BDIA

Presented by :
Haolun Yang

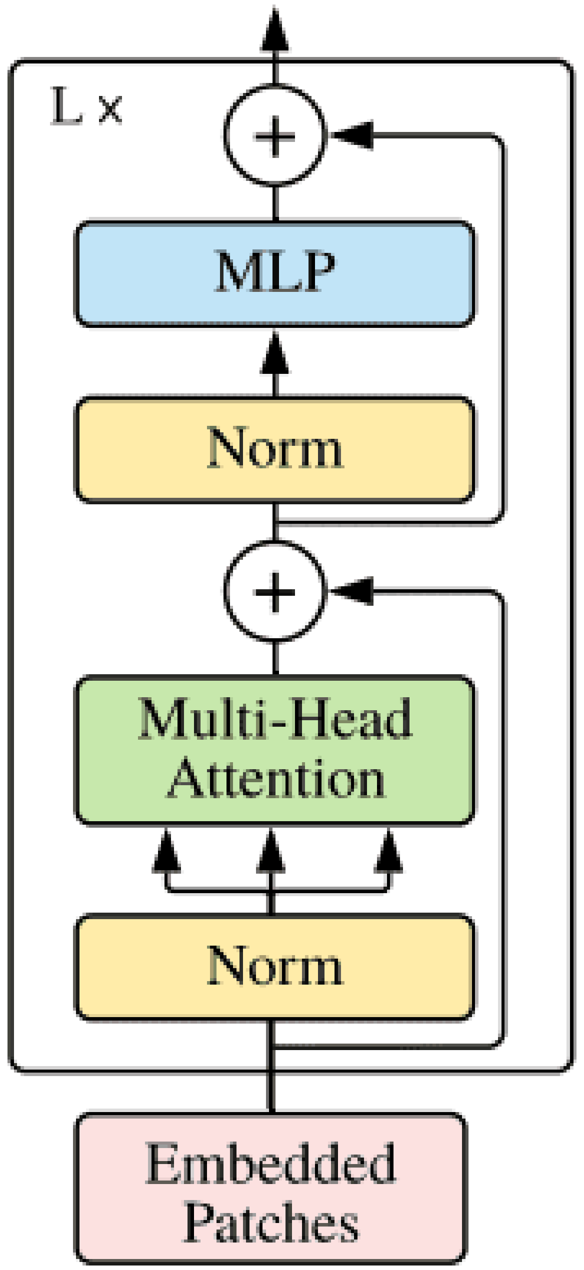
Presentation time :
6/8/2024

Topics

Background & Introduction

What is a Vision Transformer (ViT)?

A class of models that leverage self-attention mechanisms to process visual data



Advantages of ViT

* GLOBAL CONTEXTUAL UNDERSTANDING

long-range dependencies in images

better representation learning compared to CNNS

---

* SCALABILITY

Easier to scale up and receptive field size

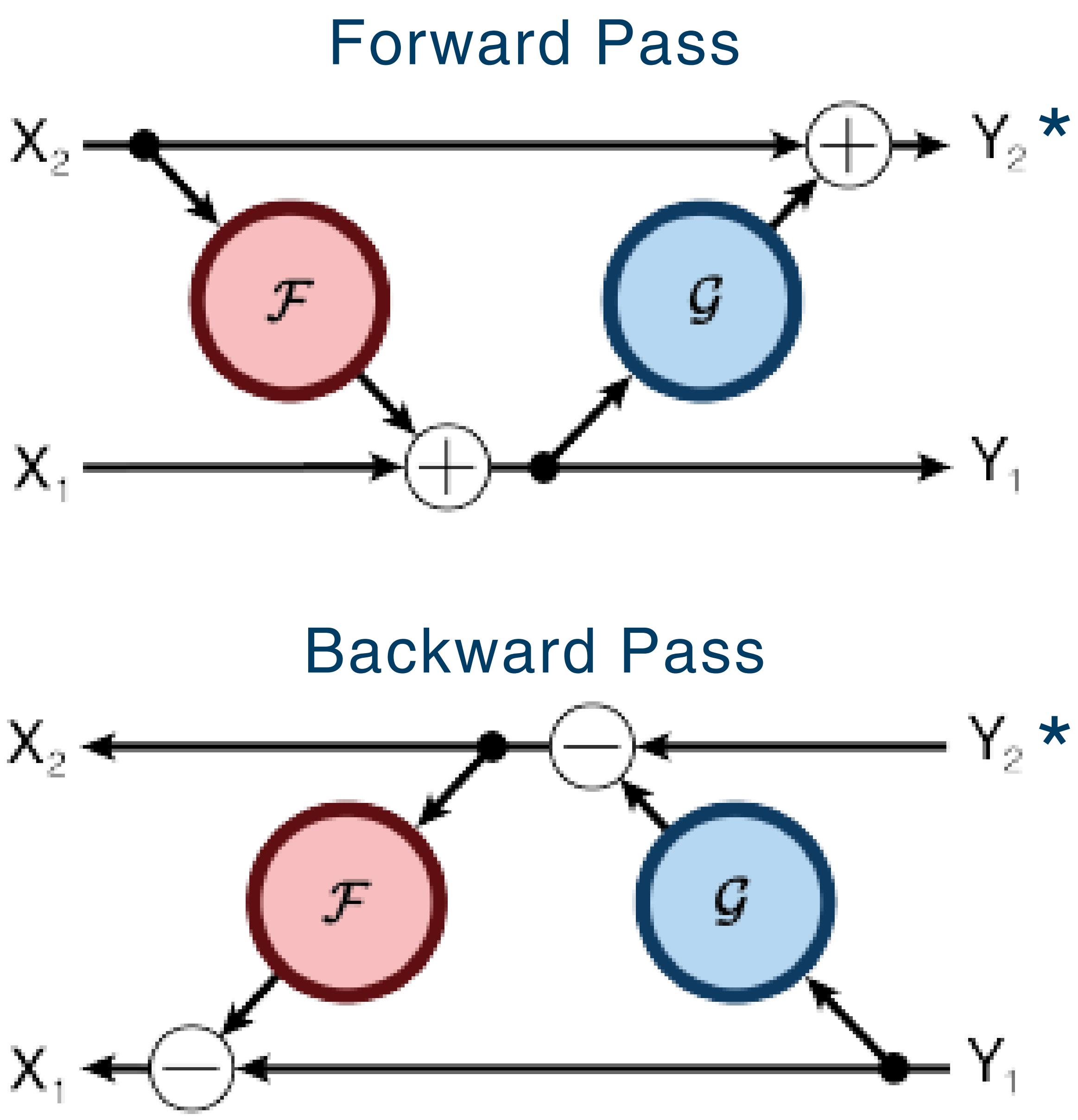
---

* TRANSFER LEARNING

Effective in leveraging pre-trained models

Background & Introduction

What are Reversible Architecture Models?



Reversible models avoid the need to store these activations, thereby reducing memory usage significantly.

* Source: Aidan N. Gomez, Mengye Ren, Raquel Urtasun, Roger B. Grosse, "The Reversible Residual Network: Backpropagation Without Storing Activations," 2017. Available at: https://arxiv.org/abs/1707.04585

Advantages of Using Reversiable Architecture

* **REDUCED MEMORY FOOTPRINT**

  Enables training deeper networks with the same amount of memory.

* **PARTICULARLY USEFUL FOR LARGE-SCALE VISION TASKS**

  Easier to trace the flow of data through the network.

* **IMPROVED SCALABILITY**

  Allows more layers or parameters to be added, thus supporting more complex models.

Background & Introduction

An optimization technique in deep learning

What is Momentum and How Momentum Works?

Update Formula:

$$v_t = \beta v_{t-1} + (1-\beta)\nabla L(\theta_t)$$

$$\theta_{t+1} = \theta_t - \alpha v_t$$

- v_t is the momentum term.
- β is the momentum hyperparameter (typically around 0.9).
- ∇L(θ_t) is the current gradient.
- α is the learning rate.

Advantages of Using Momentum

* **ACCELERATED CONVERGENCE**

  Reaches the minimum of the objective function faster.

  Reduces oscillations in gradient descent

---

* **IMPROVED STABILITY**

  Helps prevent the optimizer from getting stuck in local minima

---

* **REDUCES RANDOMNESS**

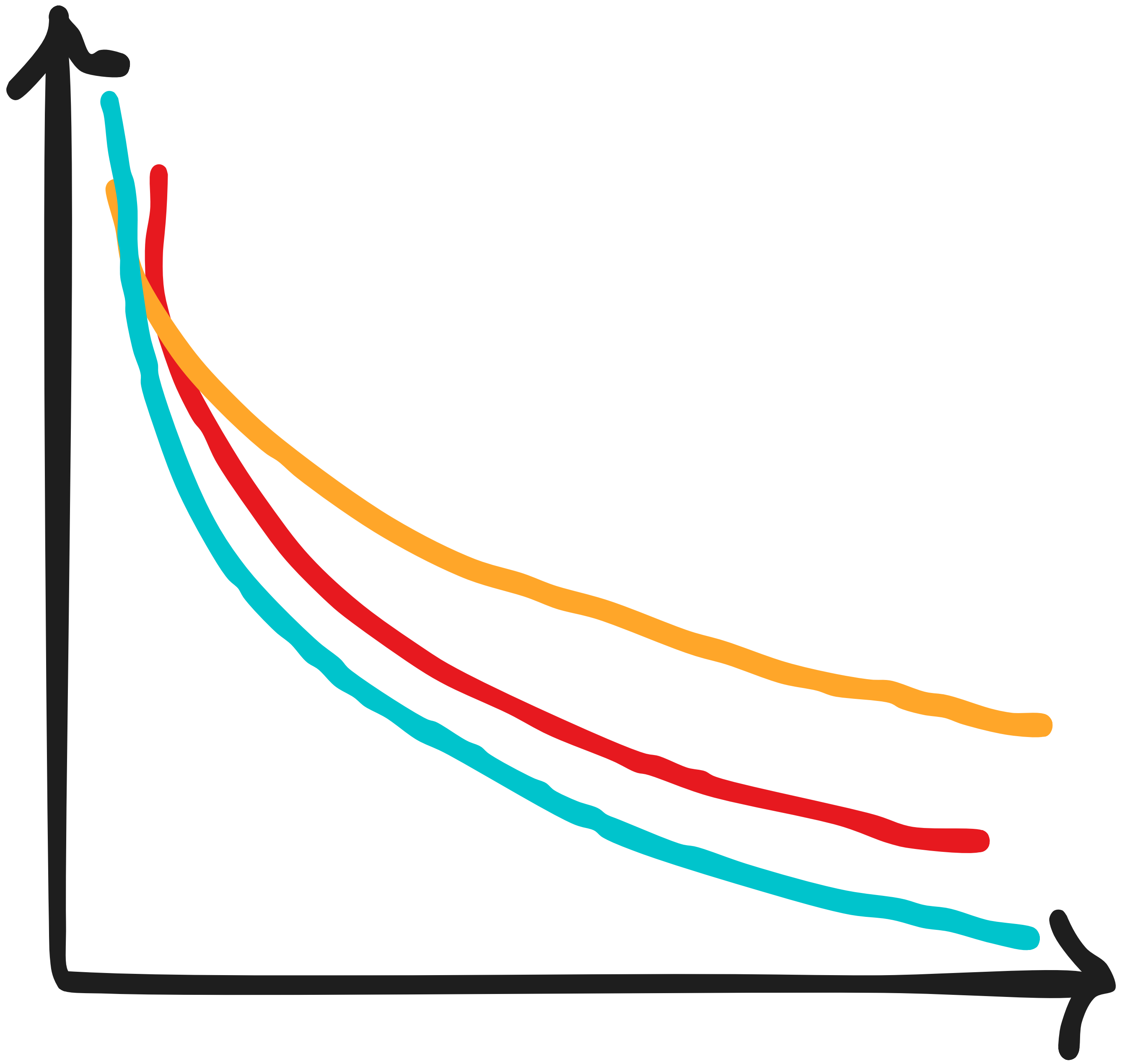
  Smoothes out the impact of noisy data and enhance overall optimization effectiveness

Research Questions

- **How can we find the optimal Vision Transformer model that balances performance and resource consumption?**

- **What techniques can enhance the performance of Reversible Vision Transformers?**

Objectives



*** Evaluate and compare the performance and resource consumption of ViT-small, Reversible ViT-small and ViT-small with BDIA models.**

- Performance Metrics: Accuracy, Speed.
- Resource Consumption: Memory, GFLOPs, Parameters.

*** Explore the application of momentum in Reversible ViT.**

- Try to analyse the impact of momentum on convergence speed and stability.

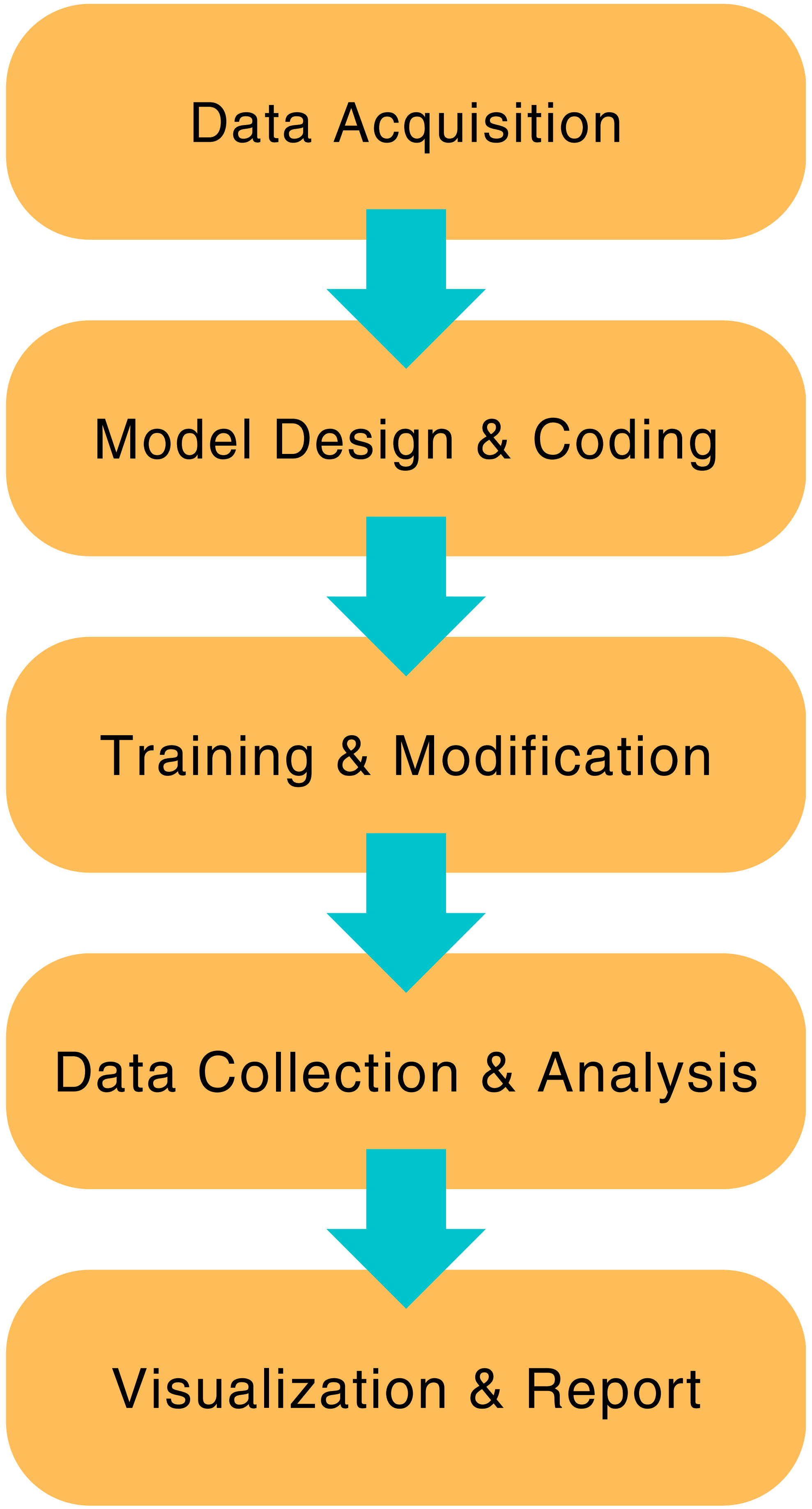
Methodology

- Use the Same Dataset CIFAR-10

- Same Base Model, Same Data Augmentation, Preprocessing and Hyperparameters.

- Train and Validate on the Same Testing Platform NVIDIA GeForce RTX 4060 Laptop GPU(8GB, CUDA).

- Testing performances of different models.



Model Explanation

* ViT small
* Reversible ViT small
* ViT small with BDIA

Vision Transformer small

- A specialized transformer model (for small datasets) designed for image classification tasks.

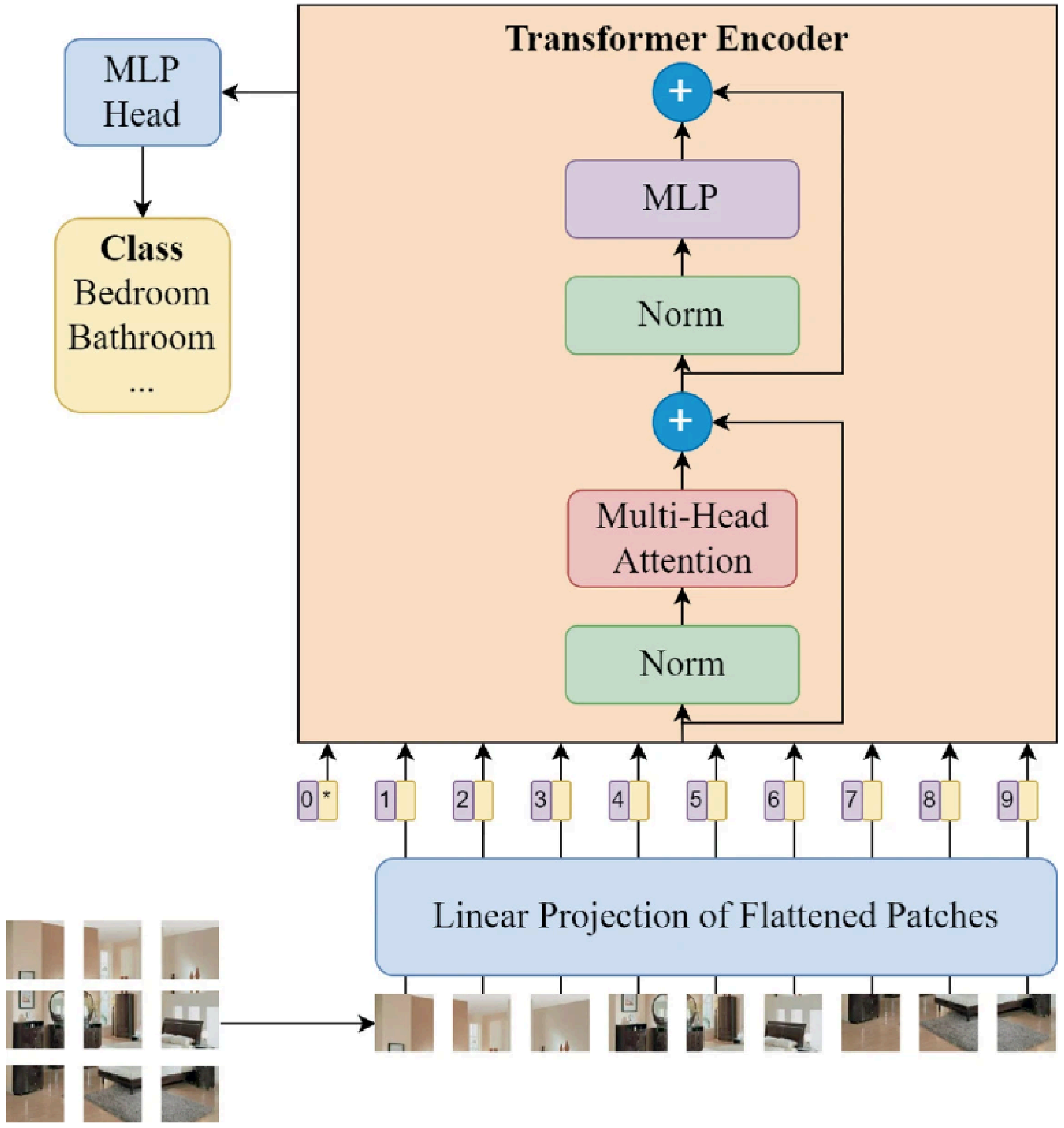
- It utilizes a unique approach of processing images as sequences of patches instead of relying on traditional convolutions like CNNs.

ViT small Architecture

* **PATCH EMBEDDING**
  Images are divided into patches, which are then embedded into vectors.

* **TRANSFORMER ENCODER**
  - Utilizes **multi-head self-attention** to analyze different aspects of the image simultaneously.
  - Includes **feed-forward layers** to enhance feature extraction.

* **MLP HEAD**
  The multi-layer perceptron complete the image classification task.



*

* Source: Veiga, B., Pinto, T., Teixeira, R., Ramos, C. (2023). Vision Transformers Applied to Indoor Room Classification. In: Moniz, N., Vale, Z., Cascalho, J., Silva, C., Sebastião, R. (eds) Progress in Artificial Intelligence. EPIA 2023. Lecture Notes in Computer Science(), vol 14116. Springer, Cham. https://doi.org/10.1007/978-3-031-49011-8_44

Reversible Vision Transformer small

A new architecture based on the Vision Transformers small

Reversibility:

- Reversible layers inside of standard blocks allow for the reconstruction of input data without the need to store all intermediate activations.
- Significantly reduces memory footprint during training.

Reversiable ViT small Architecture

Customed Reversible Block

- Including both attention and feedforward layers wrapped in a reversible framework.

FORWARD PASS:

$$y = x + \text{LSA}(\text{LayerNorm}(x))$$

$$z = y + \text{FFN}(\text{LayerNorm}(y))$$

BACKWARD PASS:

Reconstruction Process:

$$y = z - \text{FFN}(\text{LayerNorm}(y))$$

$$x = y - \text{LSA}(\text{LayerNorm}(x))$$

Backward Propagation Formulas:

$$\frac{\partial \mathcal{L}}{\partial y} = \frac{\partial \mathcal{L}}{\partial z} \cdot \left(1 + \frac{\partial \text{FFN}(\text{LayerNorm}(y))}{\partial y}\right) \qquad \frac{\partial \mathcal{L}}{\partial x} = \frac{\partial \mathcal{L}}{\partial y} \cdot \left(1 + \frac{\partial \text{LSA}(\text{LayerNorm}(x))}{\partial x}\right)$$

Vision Transformer with Bidirectional Integration Approximation (BDIA)

- A technique designed to achieve bit-level reversibility in deep learning models without changing their architectures.
- Improve the model's performance through the regularization effect of BDIA.

- Exact bit-level reversibility.
- Use of activation quantization for precise computation.
- Introducing randomness with the hyper-parameter γ.

How BDIA Works in Vision Transformer

- BDIA employs a method of approximating the forward and backward integration for each transformer block.

Original Transformer Update:

$$x_{k+1} = x_k + f_k(x_k) + g_k(x_k + f_k(x_k))$$

BDIA Update with Random Parameter γ

$$x_{k+1} = \gamma x_{k-1} + (1-\gamma)x_k + (1+\gamma)h_k(x_k)$$

where $\gamma \in \{-0.5, 0.5\}$ is randomly chosen for each training sample and transformer block.

Quantization:

$$Q_l[y] = \mathrm{round}[y/2^{-l}]2^{-l}$$

which l is the precision level.

Reversiability in ViT with BDIA

FORWARD PASS:

- Bidirectional Integration Approximation

when k = 0: $x_1 = x_0 + h_0(x_0)$

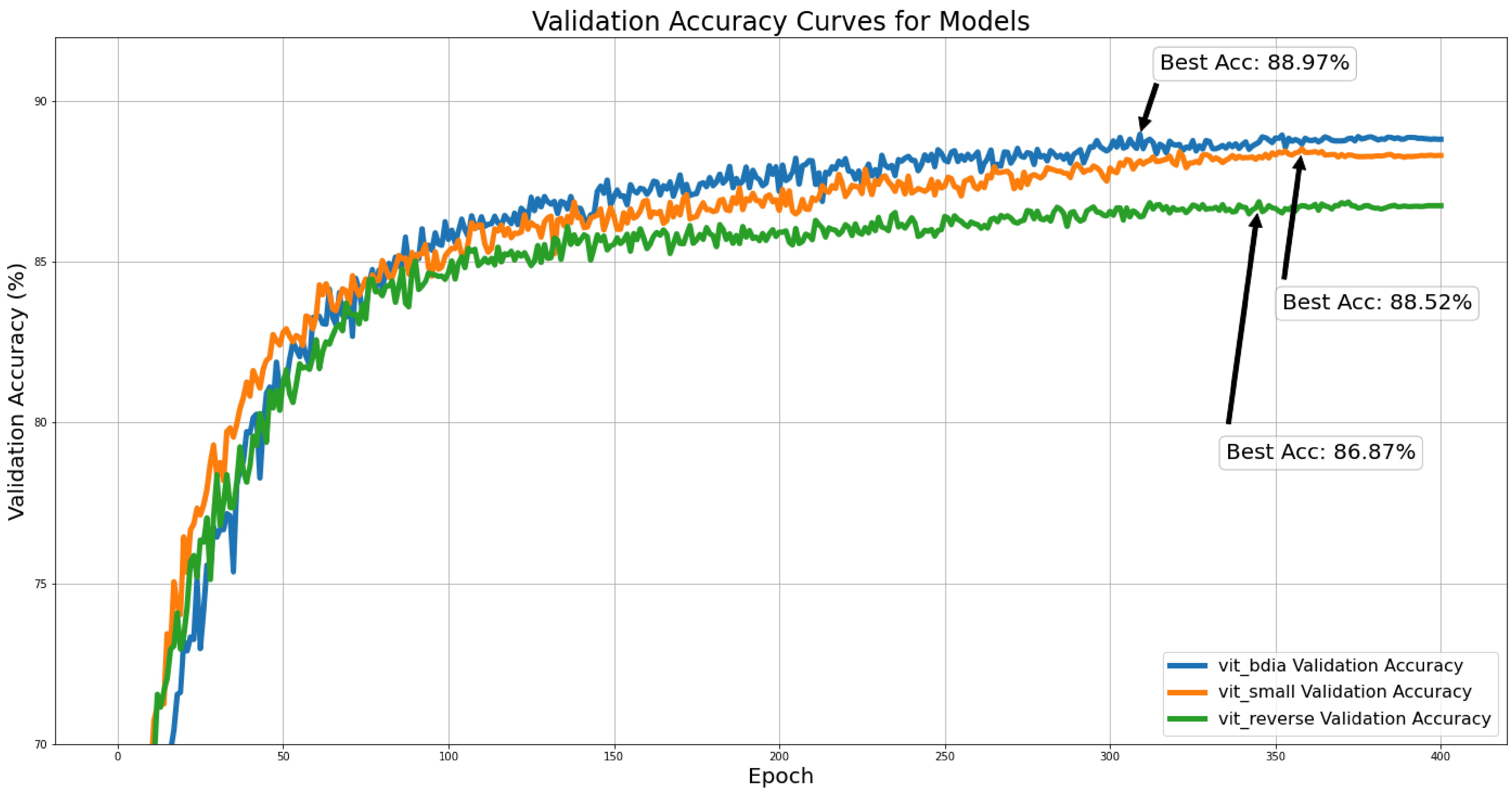
for N-1 ⩾ k > 0: $x_{k+1} = \gamma x_{k-1} + (1-\gamma)x_k + (1+\gamma)h_k(x_k)$

BACKWARD PASS:

- Exact Reversibility
- Recomputing intermediate activations on-the-fly

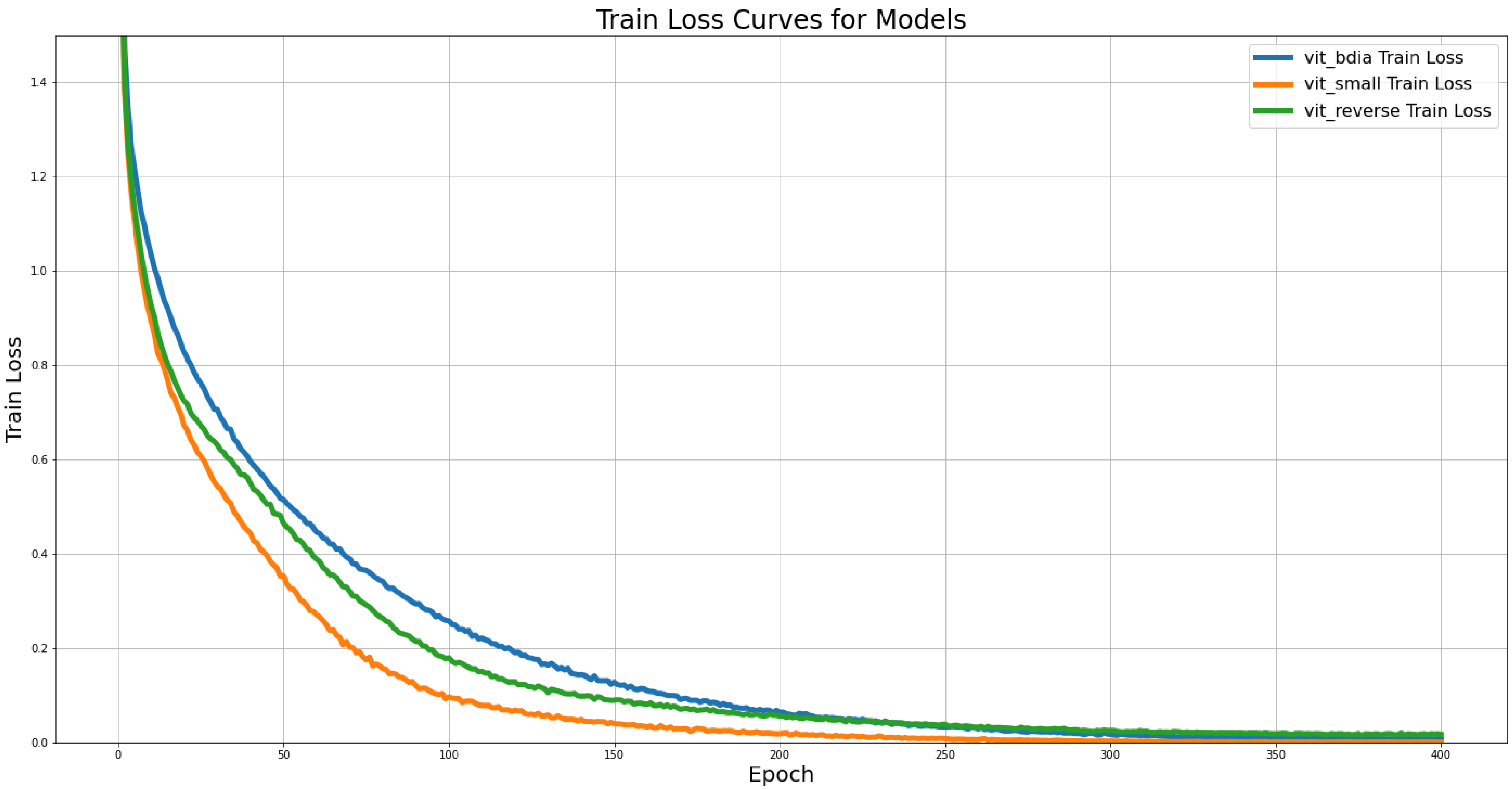
$$x_{k-1} = \frac{x_{k+1}}{\gamma} - \frac{1-\gamma}{\gamma}x_k - \frac{1+\gamma}{\gamma}h_k(x_k)$$

Results



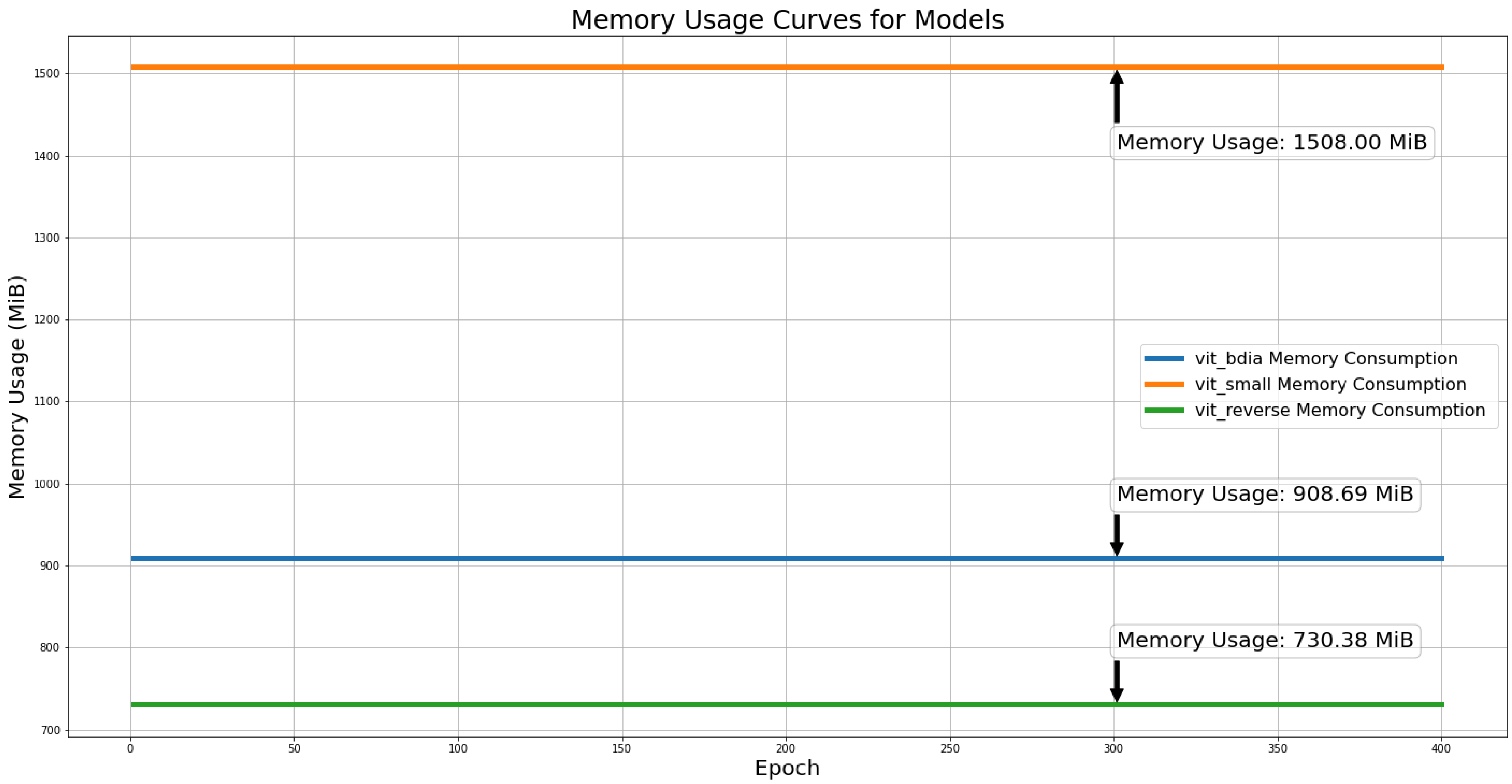
Validation Accuracy Curves for Models
Best Acc: 88.97%
Best Acc: 88.52%
Best Acc: 86.87%
90
85
80
75
70
Validation Accuracy (%)
0
50
100
150
200
250
300
350
400
Epoch
vit_bdia Validation Accuracy
vit_small Validation Accuracy
vit_reverse Validation Accuracy

Results



Train Loss Curves for Models
vit_bdia Train Loss
vit_small Train Loss
vit_reverse Train Loss
Train Loss
Epoch
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
0
50
100
150
200
250
300
350
400

Results


Memory Usage Curves for Models
Memory Usage (MiB)
1500
1400
1300
1200
1100
1000
900
800
700
0
50
100
150
200
250
300
350
400
Epoch
Memory Usage: 1508.00 MiB
Memory Usage: 908.69 MiB
Memory Usage: 730.38 MiB
vit_bdia Memory Consumption
vit_small Memory Consumption
vit_reverse Memory Consumption

Results

Model	Memory (MB/img)	GFLOPs	Param(M)
ViT-small	11.78	1.237	9.59
Reversiable ViT-small	5.71	1.237	9.59
ViT-small BDIA	7.10	1.237	9.59

Conclusion

- ViT Small:
  - Advantages: Simple and straightforward architecture.
  - Disadvantages: Higher memory usage compared to the other two models.

- ViT Reverse:
  - Advantages: Best in memory efficiency, suitable for environments with limited resources.
  - Disadvantages: Higher training complexity due to the intricate backpropagation mechanism and lower accuracy.

- ViT BDIA:
  - Provides a balance between performance and resource consumption.

NEXT STEP

- Momentum Reversible ViT Model:
  - Integrate momentum into the reversible ViT model to accelerate convergence, enhance training stability, and improve overall performance.
  - Address issues that arise when implementing momentum mechanisms, such as vanishing or exploding gradients and instabilities during training.

- Explore and Compare Other Reversible Models

- Practical Applications

Thank You!

For questions and concerns, feel free to get in touch.

Presented by :
Haolun Yang

Email:
hy383@exeter.ac.uk

Presentation Video Link

OneDrive:

https://universityofexeteruk-my.sharepoint.com/:v:/g/personal/hy383_exeter_ac_uk/EYDFIL38EyFGlxIhjDzB4goBMSh8v1UEDkt9QVd3iQN8yg?nav=eyJyZWZlcnJhbEluZm8iOnsicmVmZXJyYWxBcHAiOiJPbmVEcml2ZUZvckJ1c2luZXNzIiwicmVmZXJyYWxBcHBQbGF0Zm9ybSI6IldlYiIsInJlZmVycmFsTW9kZSI6InZpZXciLCJyZWZlcnJhbFZpZXciOiJNeUZpbGVzTGlua0NvcHkifX0&e=7NzIbl

Youtube:

https://youtu.be/kbNW3p_PTJ4