NEURAL INFORMATION PROCESSING SYSTEMS 2024

HUAZHONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

MoE Jetpack: From Dense Checkpoints to Adaptive Mixture of Experts for Vision Tasks

Xingkui Zhu*, Yiran Guan*, Dingkang Liang, Yuchao Chen,

Yuliang Liu, Xiang Bai

Huazhong University of Science and Technology

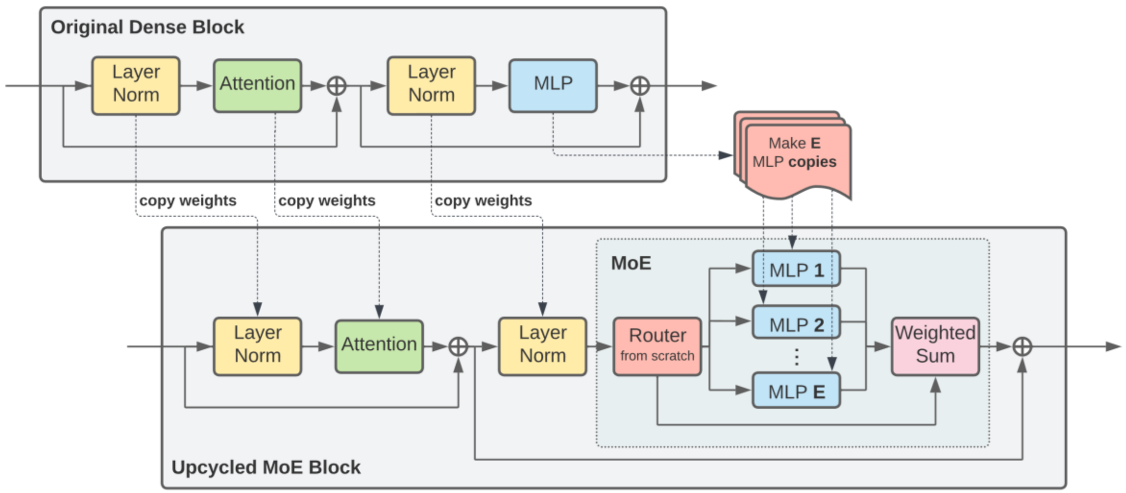
VLR Group

Background

What is Mixture of Experts (MoE)?

MoE architecture comprises:

- Densely activated layers
- Routers + Sparsely activated MoE layers



Why MoE?

Advantages:

- **Scalability:** Allows model scaling with minimal increase in inference cost (FLOPs).
- **Efficiency:** Achieves faster training and inference compared to dense models with similar parameter counts.
- **Performance:** Delivers improved performance at similar inference speeds to dense models.

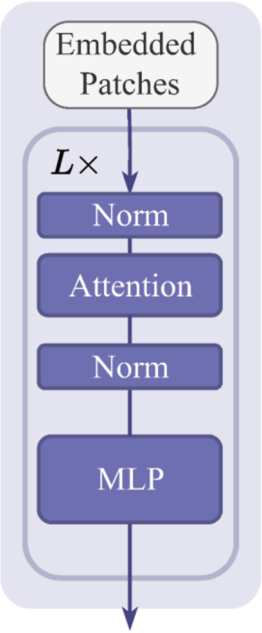
[1] Komatsuzaki, Aran, et al. "Sparse Upcycling: Training Mixture-of-Experts from Dense Checkpoints." *The Eleventh International Conference on Learning Representations*.

Motivation

Abundant
vision dense checkpoints

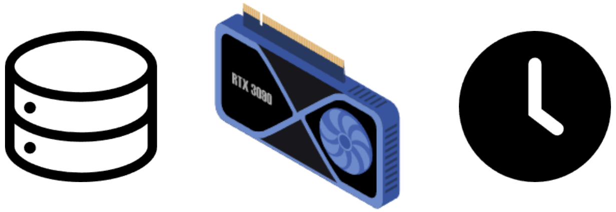
Hugging Face

Pre-trained
Dense Model

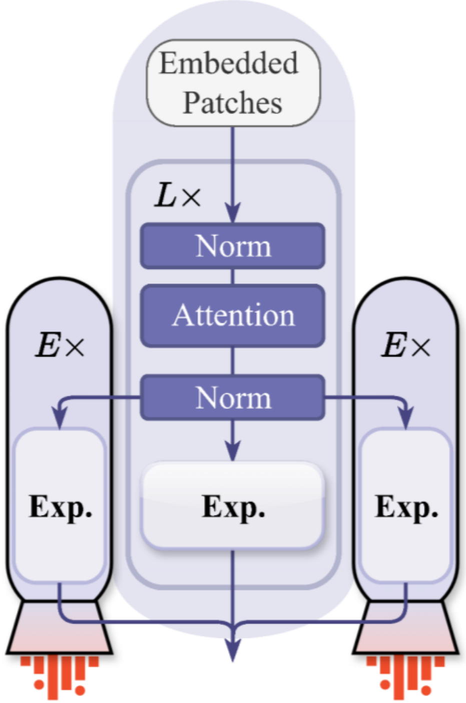


Scarce
MoE checkpoints

Pre-training



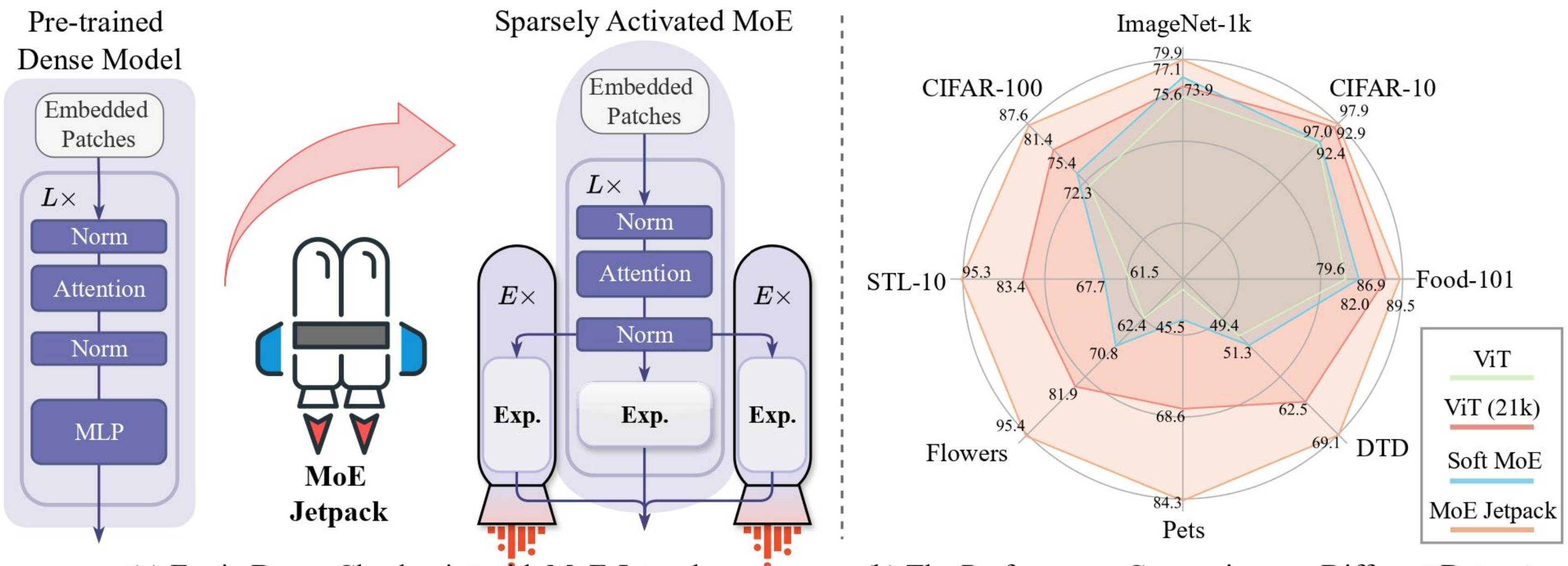
Sparsely Activated MoE



How to maximize the use of ***dense checkpoints*** *to enhance*
the accuracy and convergence speed of ***MoE models*** *during fine-tuning?*

MoE Jetpack

Our MoE Jetpack leverages dense checkpoints to **bypass the MoE pre-training phase**, capitalizing on sunk pre-training costs to achieve **faster convergence** and **enhanced performance**.



(a) Equip Dense Checkpoint with MoE Jetpack

(b) The Performance Comparison on Different Datasets

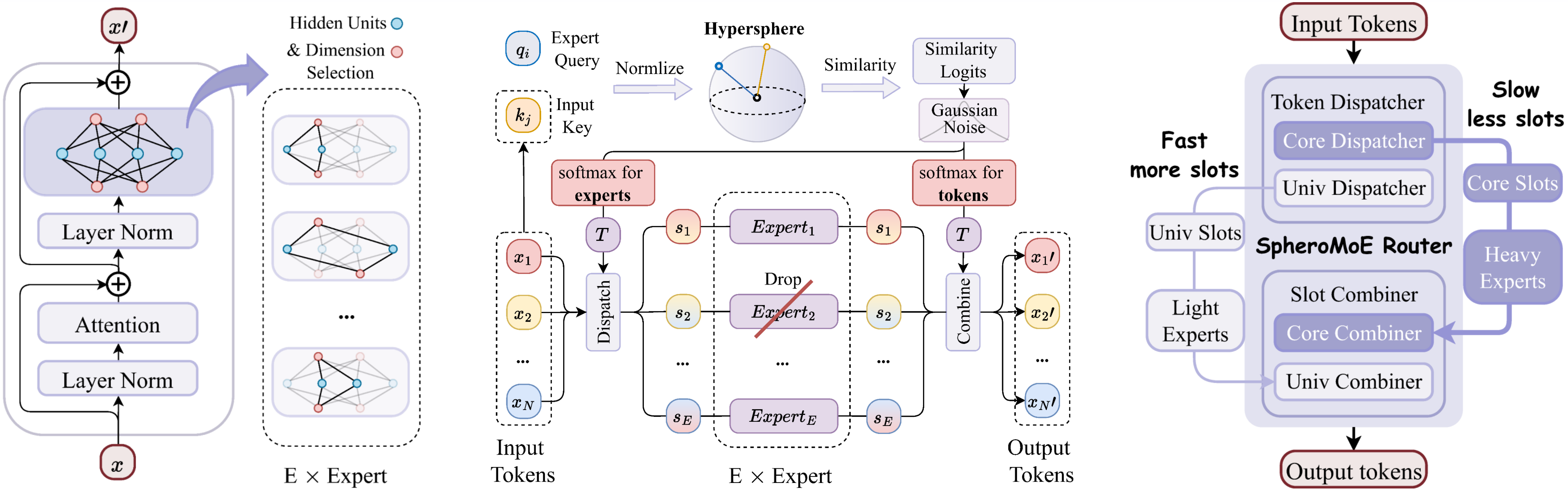
Highlights:

- **Stronger performance.**
- **Faster Convergence.**
- **Robust generalization.**
- **Running Efficiency.**

Method Overview

MoE Jetpack is a framework that fine-tunes pre-trained dense models into Mixture of Experts with:

a) Checkpoint Recycling and **b) SpheroMoE Layers** which contain **c) Adaptive Dual-path**.



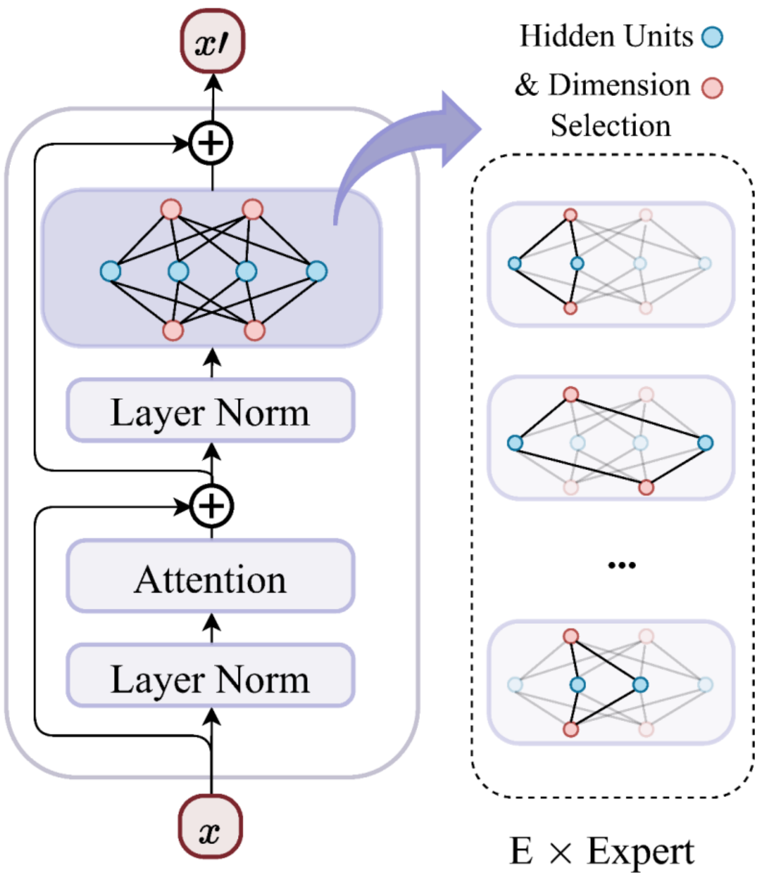
(a) Checkpoint Recycling (b) SpheroMoE Layer (c) Adaptive Dual-path

Method

1 Checkpoints recycling

- Random
- Uniform
- **L2**
- Co-Activation Graph Split

Unlike existing methods that simply replicate the Feed-Forward Network (FFN) to construct MoE, our approach employs **importance sampling** to select **diverse experts** with **varying sizes**, enabling more effective MoE weight initialization.

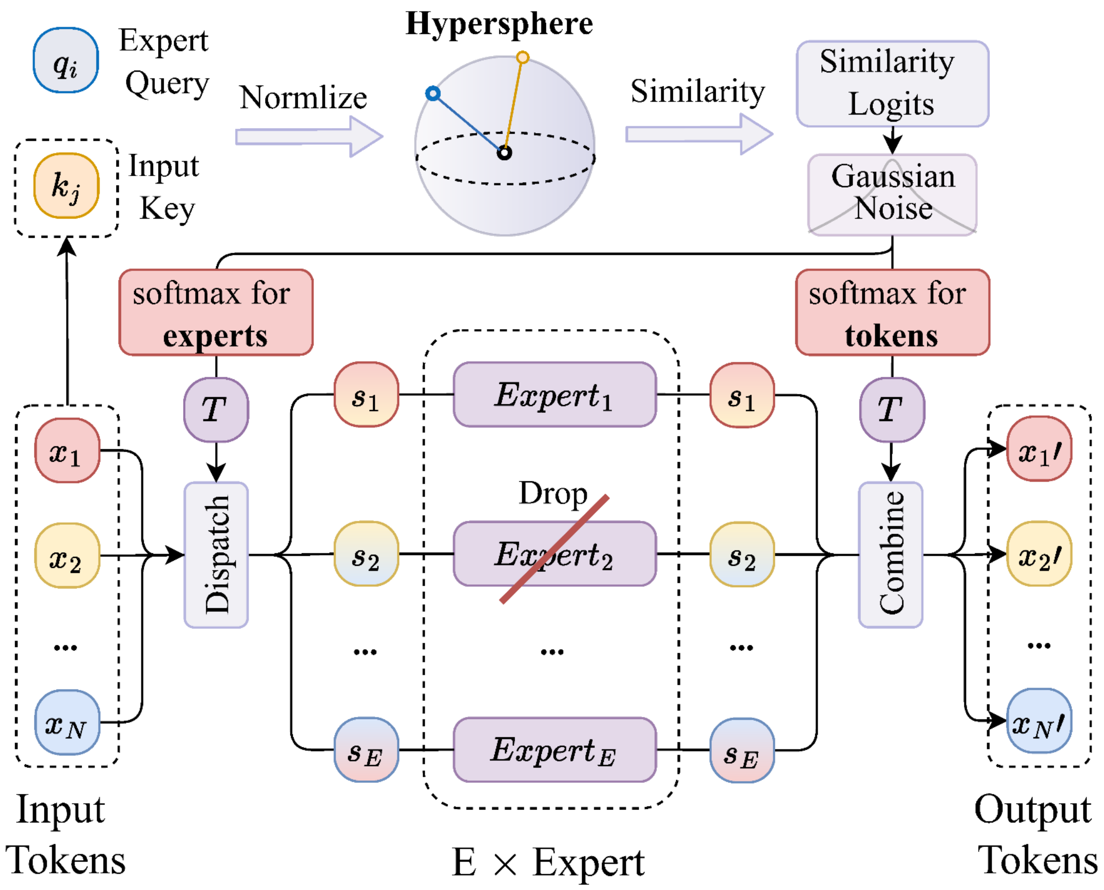


Method

2 SpheroMoE Layers

- Hypersphere projection
- Learnable SoftMax temperature
- MoE regularization

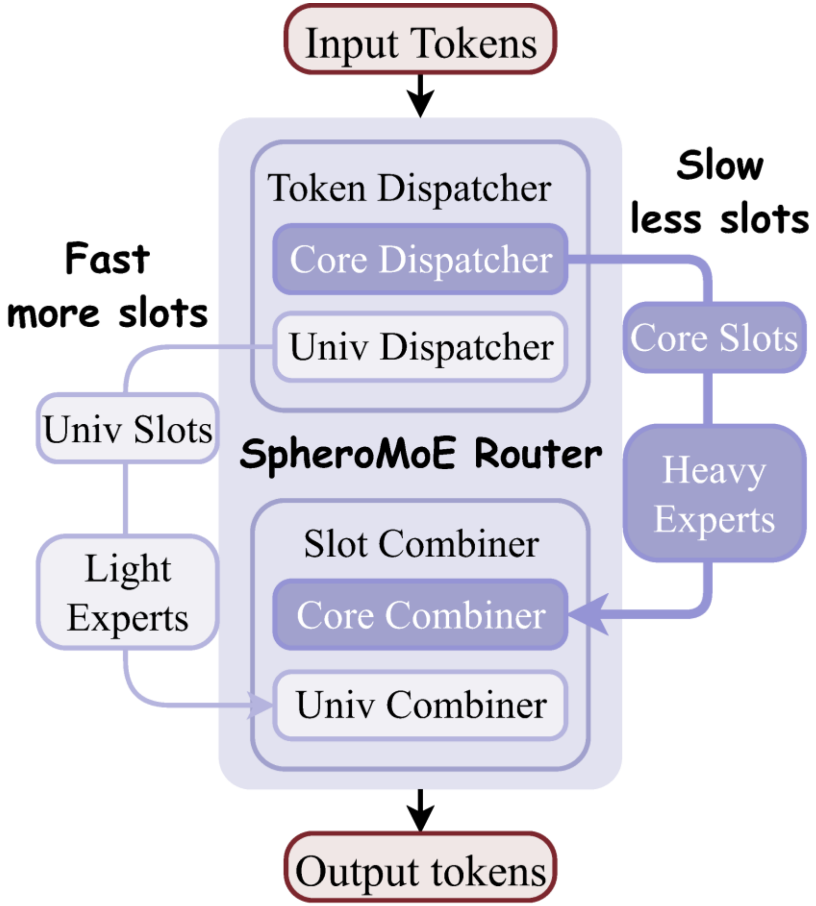
SpheroMoE Layers employ cross-attention to reorganize input tokens into slots for expert processing. It's a continuous, fully differentiable routing mechanism based on Soft MoE.



Method

3 Adaptive Dual-path

The dual-path structure directs **important tokens** to **large core experts** and routes less critical tokens to smaller, numerous experts, enhancing efficiency without losing performance.



Experiments

The MoE Jetpack, benefiting from the pre-trained knowledge embedded in dense checkpoints, **consistently surpasses** the performance of both Soft MoE models trained from scratch and dense models with ImageNet-21K initialization.

Performance comparison on visual recognition tasks.

Dataset (↓)	Dense	Dense (21k)	Soft MoE [6]	MoE Jetpack
ImgNet-1k	73.9	75.6	77.1	79.9 (+2.8)
Food-101	79.6	86.9	82.0	89.5 (+7.5)
CIFAR-10	92.4	97.0	92.9	97.9 (+5.0)
CIFAR-100	72.3	81.4	75.9	88.4 (+12.5)
STL-10	61.5	83.4	67.7	95.3 (+27.6)
Flowers	62.4	81.9	70.8	95.4 (+24.6)
Pets	25.0	68.6	45.5	84.3 (+38.8)
DTD	49.4	62.5	51.3	69.1 (+17.8)

ViT

Dataset (↓)	Dense	Dense (21k)	Soft MoE [6]	MoE Jetpack
ImgNet-1k	76.1	76.4	79.1	80.5 (+1.4)
Food-101	86.9	89.0	88.7	90.7 (+2.0)
CIFAR-10	96.6	97.4	97.3	98.2 (+0.9)
CIFAR-100	81.4	84.4	82.8	88.5 (+5.7)
STL-10	81.4	92.3	79.4	98.7 (+19.3)
Flowers	80.3	94.5	83.3	98.6 (+15.3)
Pets	72.9	87.3	77.4	94.9 (+17.5)
DTD	63.7	68.8	64.7	79.5 (+14.8)

ConvNeXt

Experiments

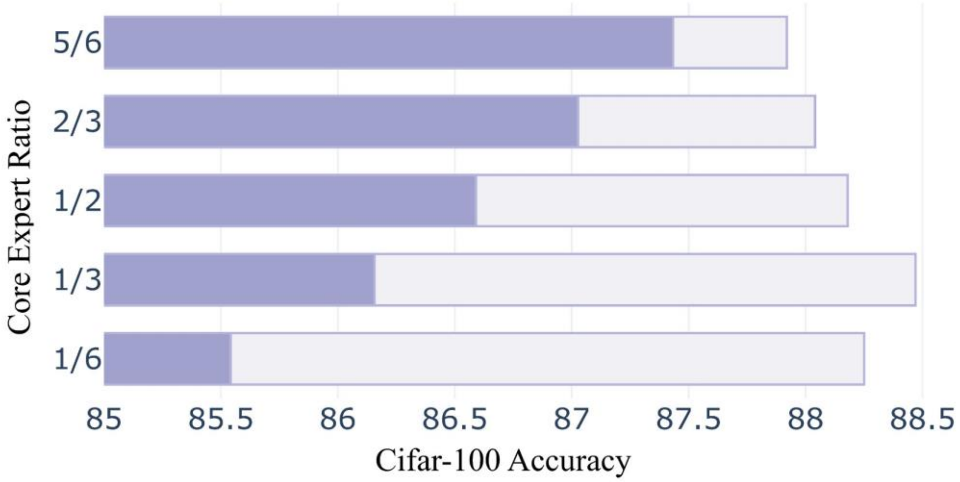
Ablation study on MoE Jetpack Components.

Soft MoE [6]	Checkpoints Recycling	SpheroMoE	ImageNet	CIFAR-100	Flowers	Mean Acc.
Baseline ViT-T			73.9	72.3	62.4	69.5
✓			77.1	75.9	70.8	74.6 (+5.1)
✓	✓		78.4	84.7	91.2	84.8 (+15.3)
	✓	✓	79.9	88.4	95.4	**87.9** (+18.4)

Ablation study on Checkpoint Recycling Methods.

Method	Construction	ImageNet
Sparse Upcycling [16]	Copy	79.1
Checkpoint Recycling	Random Sampling	79.5
	Uniform Selection	79.6
	Graph Partitioning	79.8
	Importance-based Sampling	**79.9**

Ablation study on Core Expert Ratio.

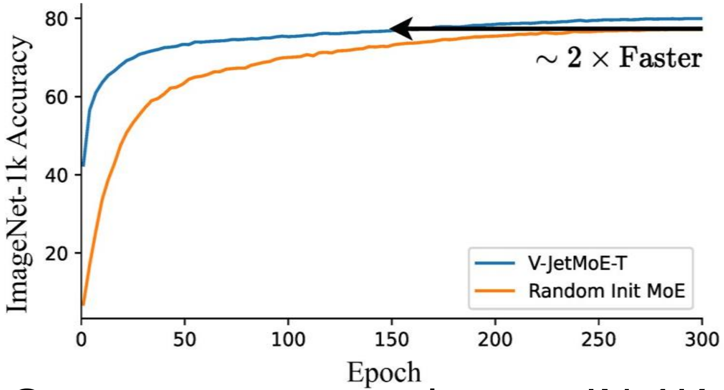


Experiments

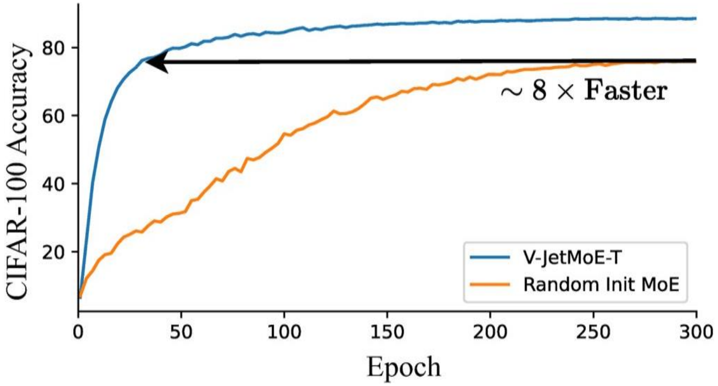
Comparison of Model Variants with Different Configurations

model	Weight Init.	MoE Layers	Expert Number	Param (M)	FLOPs (G)	CIFAR-100	ImageNet
ViT-T	-	-	-	6	1.1	72.3	73.9
Soft MoE-T [6]	-	7:12	197	354	1.2	75.9	77.1
Soft MoE-S [6]	-	7:12	197	1412	4.5	77.5	80.3
ViT-T	✓	-	-	6	1.1	81.4	75.5
V-JetMoE-T	✓	11:12	core: 98, univ: 196	92	1.1	87.4	-
V-JetMoE-T	✓	9:12	core: 98, univ: 196	179	1.1	87.8	-
V-JetMoE-T	✓	5:12	core: 98, univ: 196	352	1.2	86.7	-
V-JetMoE-T	✓	7:12	core: 32, univ: 64	89	0.8	87.8	-
V-JetMoE-T	✓	7:12	core: 64, univ: 128	175	1.0	88.0	-
V-JetMoE-T	✓	7:12	core: 98, univ: 196	265	1.1	<u>88.4</u>	<u>79.9</u>
V-JetMoE-S	✓	7:12	core: 98, univ: 196	1058	4.3	**89.9**	**82.4**

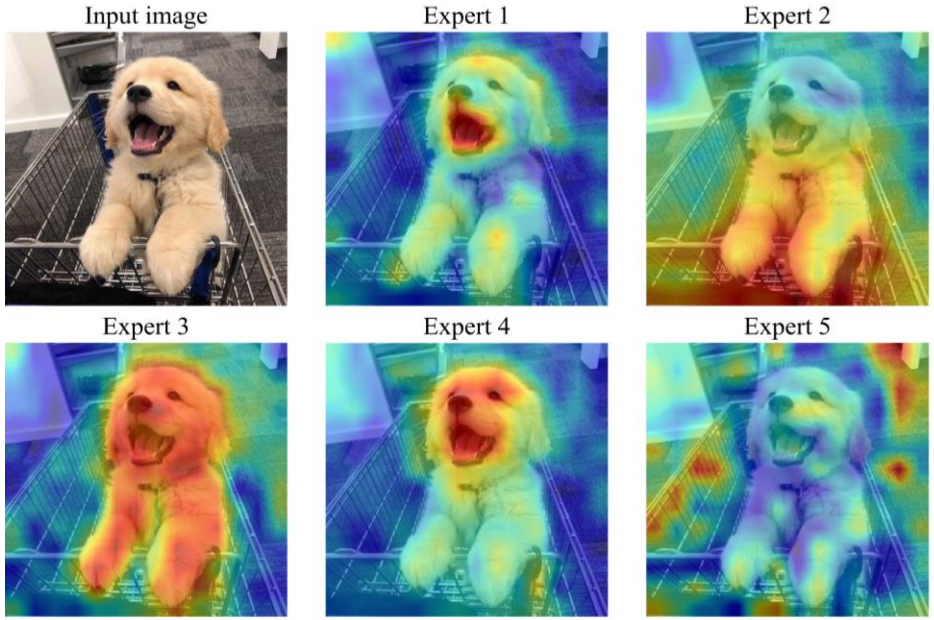
Experiments



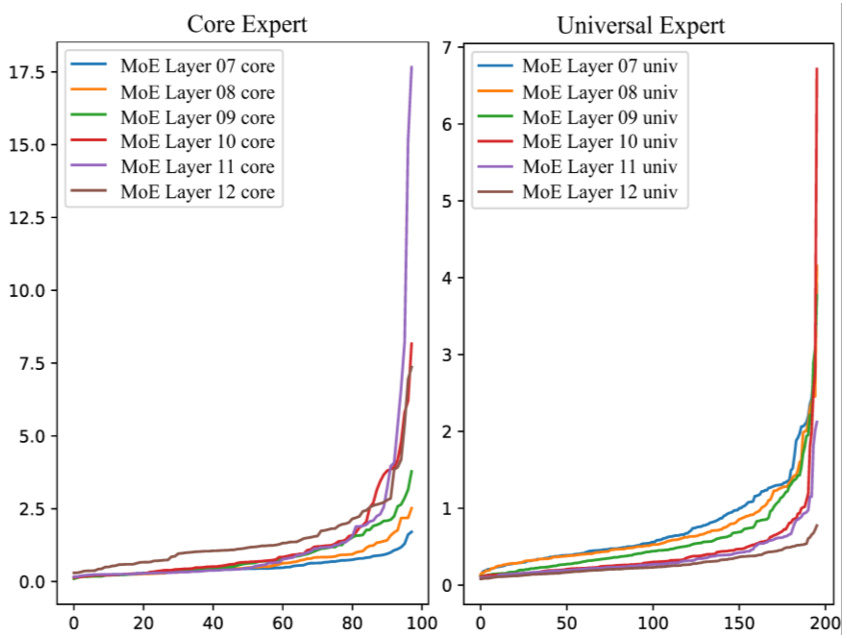
Convergence speed up on IN-1K.



Convergence speed up on CIFAR-100.



Attention Map for Experts.



Contribution for Experts.

Conclusion

MoE Jetpack: From Dense Checkpoints to Adaptive Mixture of Experts for Vision Tasks

- **Checkpoint recycling**: Pioneers the sampling of dense checkpoints to initialize MoE experts, enhancing initialization flexibility, diversifying experts, and eliminating thecomputational burden of MoE pre-training.
- **SpheroMoE layer**: Optimized for fine-tuning dense checkpoints into MoE architectures, alleviating optimization challenges, and preventing the over-specialization of experts.

NEURAL INFORMATION PROCESSING SYSTEMS

THANK YOU

Code & Models

Xingkui Zhu

https://github.com/Adlith/MoE-Jetpack