Robustifying Routers Against Input Perturbations for Sparse Mixture-of-Experts Vision Transformers

OJSP ICASSP 2025

Masahiro Kada[1] Ryota Yoshihashi[1] Satoshi Ikehata[1, 2] Rei Kawakami[1] Ikuro Sato[1, 3]

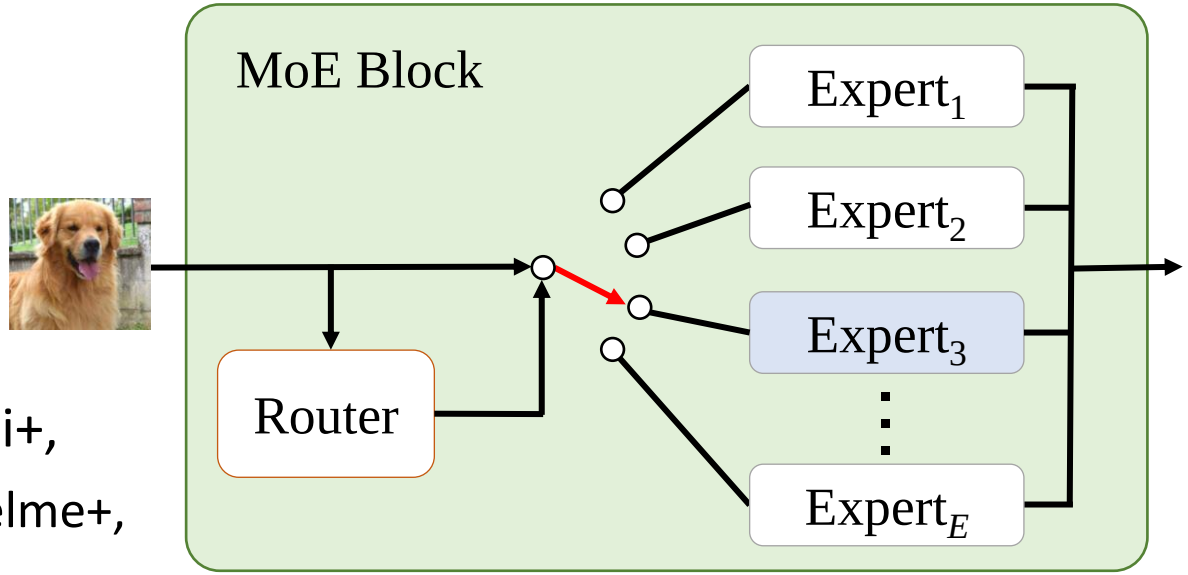
[1] Institute of Science Tokyo, Japan [2] National Institute of Informatics, Japan [3] Denso IT Laboratory, Japan

Sparse Mixture of Experts (MoE) [R. Jacobs+, Neural Comput 1991]

A sparse neural network that assigns inputs to experts using a router.

Pros

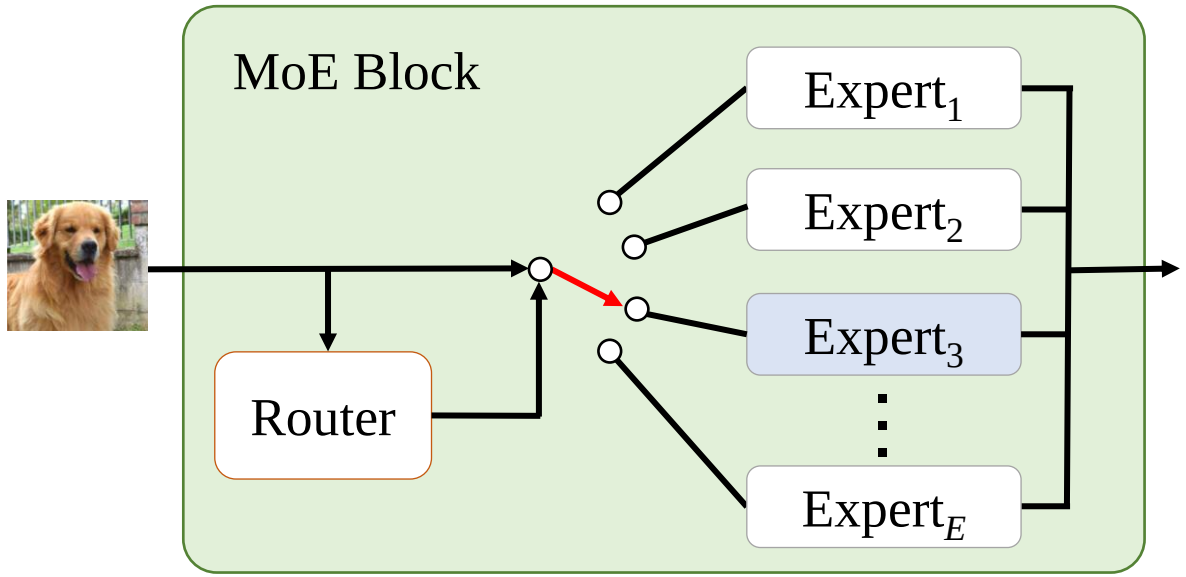
Improved model expressiveness with almost no change in calculation time.

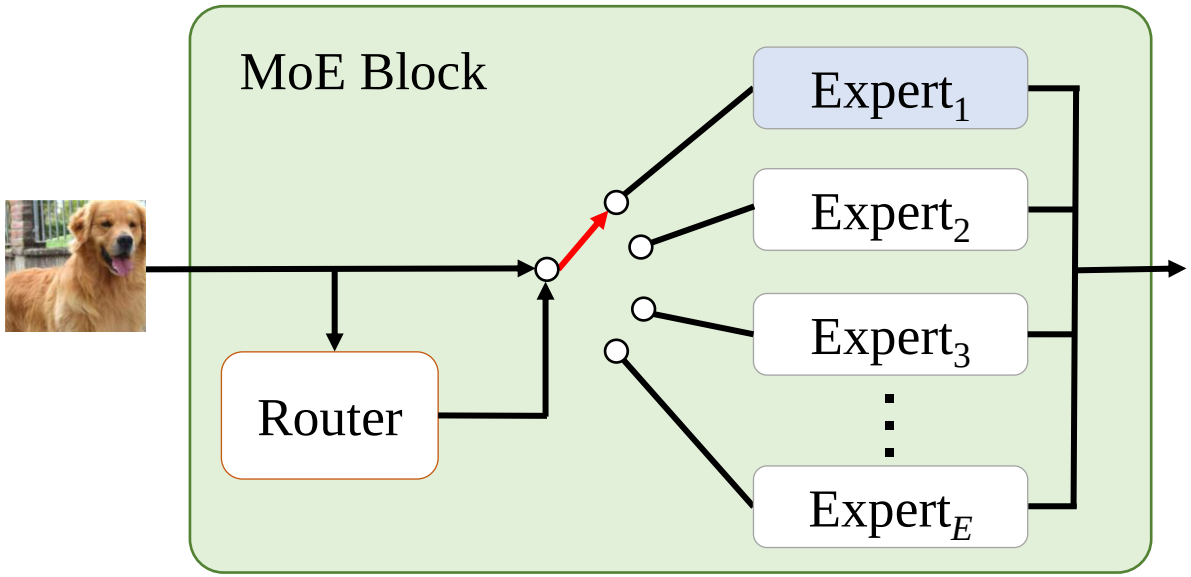


→ Used in LLM [N. Shazeer+, ICLR2017][D. Dai+, arXiv2401.06066] and vision models [C. Riquelme+, NeurIPS2021]

Introduction | Cons of Sparse Mixture of Experts

Small input perturbation may change the expert selected.

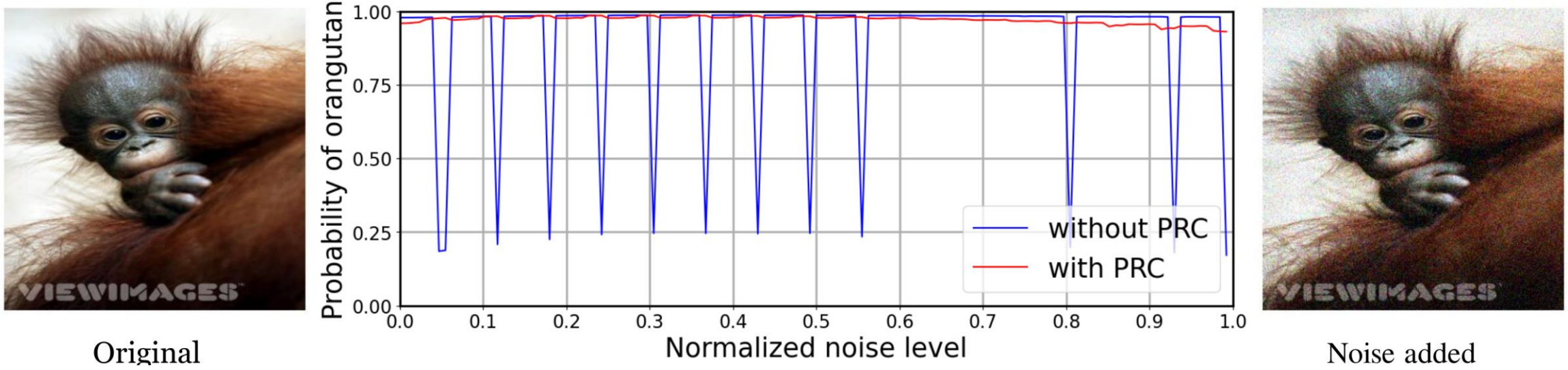




Introduction | Cons of Sparse Mixture of Experts

Small input perturbation may change the expert selected.

→ Causing the network output to change discretely and become unstable.

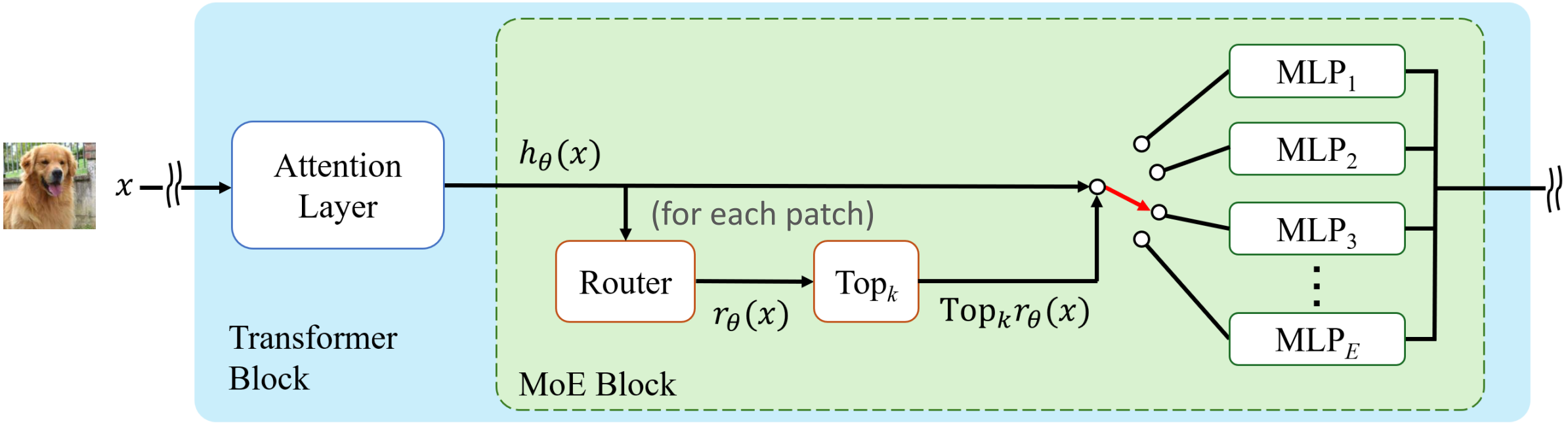


Original

Noise added

Vision Mixture of Experts (V-MoE)

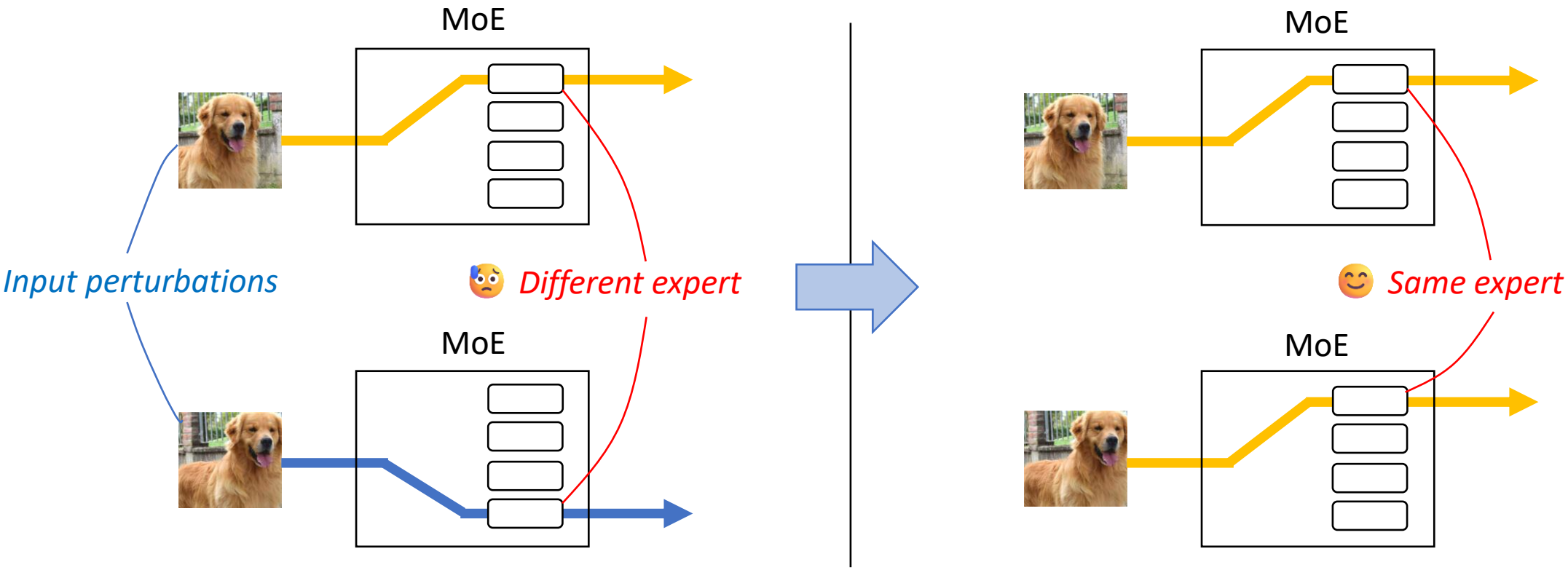
- Replacing the MLP layer of the Vision Transformer with an MoE layer.
- Routing for each image patch.



(Figure drawn by us)

Research Goal

To make the router robust against input perturbations.



Proposed Method | Pairwise Router Consistency (PRC)

The routing function $r_\theta(x)$ preferably satisfies the following conditions:

1. Be robust to data augmentation.

2. Uniformly select experts across a dataset.

3. Return an output that is close to one-hot.

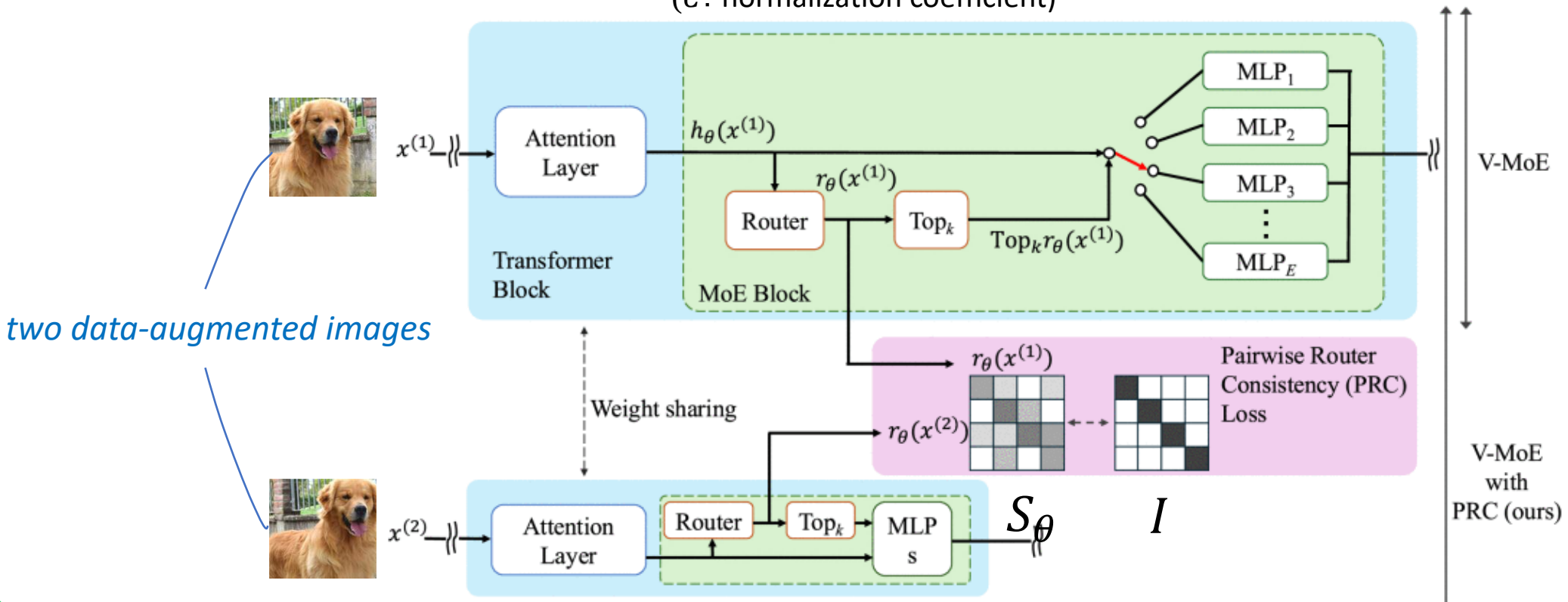
Proposed Method | Pairwise Router Consistency (PRC)

Regularization function L_θ^{PRC} promoting three requirements.

$$L_\theta^{\mathrm{PRC}} = \| S_\theta - I \|_2^2$$

$$S_\theta = C \sum_{x \in X} r_\theta\left(x^{(1)}\right) r_\theta\left(x^{(2)}\right)^T$$

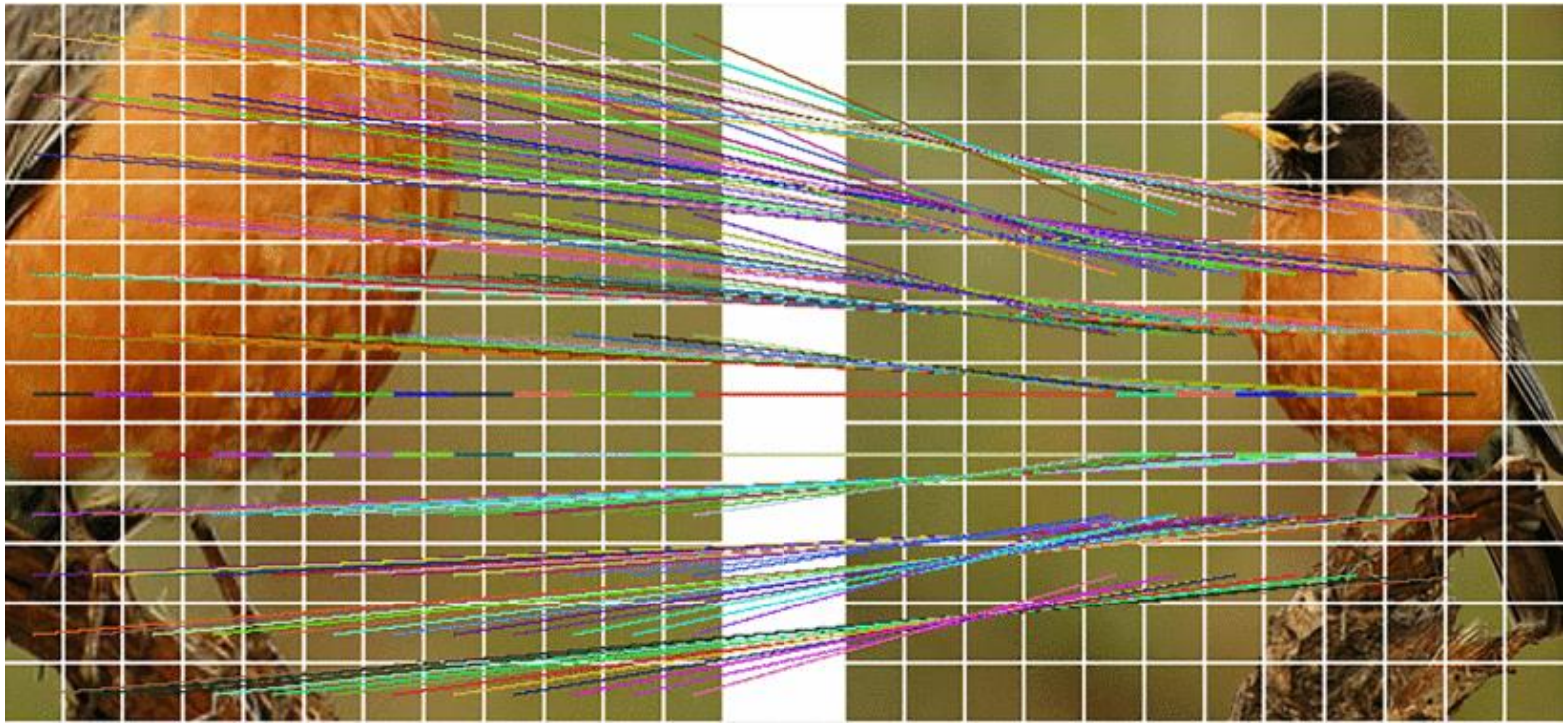
(C: normalization coefficient)



Proposed Method | Pairwise Router Consistency (PRC)

The proposed method perform routing on a patch-by-patch basis. Similar implementation is adopted in [C. Riquelme+, NeurIPS2021].

→ Corresponding patches are associated between two data-augmeneted images, and we impose PRC loss in each pair independently.



Experiments Settings

We compare the image classification accuracy of V-MoE-S and V-MoE-S with PRC to demonstrate the efficiency of our PRC regularizer.

Number of Transformer blocks	8
Layer number of MoE	6, 8
Pre-training dataset	ImageNet-1K
Fine-tuning datasets	ImageNet-1K, CIFAR-10, CIFAR-100 and Oxford Flowrs-102
Hidden layer size	512
Patch image size	32 x 32
Input image size	224 (pre-training) / 384 (fine-tuning)
Data augmentation	RandAugment (pre-training) / Random Crop, Flip (fine-tuning)
Optimizer	Adam (pre-training) / Momentum SGD (fine-tuning)
Learning rate schedule	Cosine decay with warmup

Evaluation | Image Classification Accuracy

When PRC is applied, it achieves higher accuracy than V-MoE on the ImageNet-1K, CIFAR, and Flowers datasets.

	K	ImageNet-1K	CIFAR-10	CIFAR-100	Flowers
V-MoE-S	2	75.84%	95.20%	81.38%	89.30%
V-MoE-S w/ PRC	2	**76.27%**	**95.36%**	**82.27%**	**90.18%**
V-MoE-S	1	75.23%	94.81%	81.18%	90.21%
V-MoE-S w/ PRC	1	**75.92%**	**95.12%**	**82.12%**	**91.24%**

K is the number of experts to be selected.

Evaluation | Image Classification Accuracy

When PRC is applied, it achieves higher accuracy than V-MoE on the ImageNet-1K, CIFAR, and Flowers datasets.

	K	ImageNet-1K	CIFAR-10	CIFAR-100	Flowers
V-MoE-S	2	75.84%	95.20%	81.38%	89.30%
V-MoE-S w/ PRC	2	76.27%	95.36%	82.27%	90.18%
V-MoE-S	1	75.23%	94.81%	81.18%	90.21%
V-MoE-S w/ PRC	1	**75.92%**	**95.12%**	**82.12%**	**91.24%**

K is the number of experts to be selected.

PRC with K=1 achieves accuracy comparable to V-MoE with K=2, while halving the computation of expert layers.

Analyses | Router Robustness Against Data Augmentation

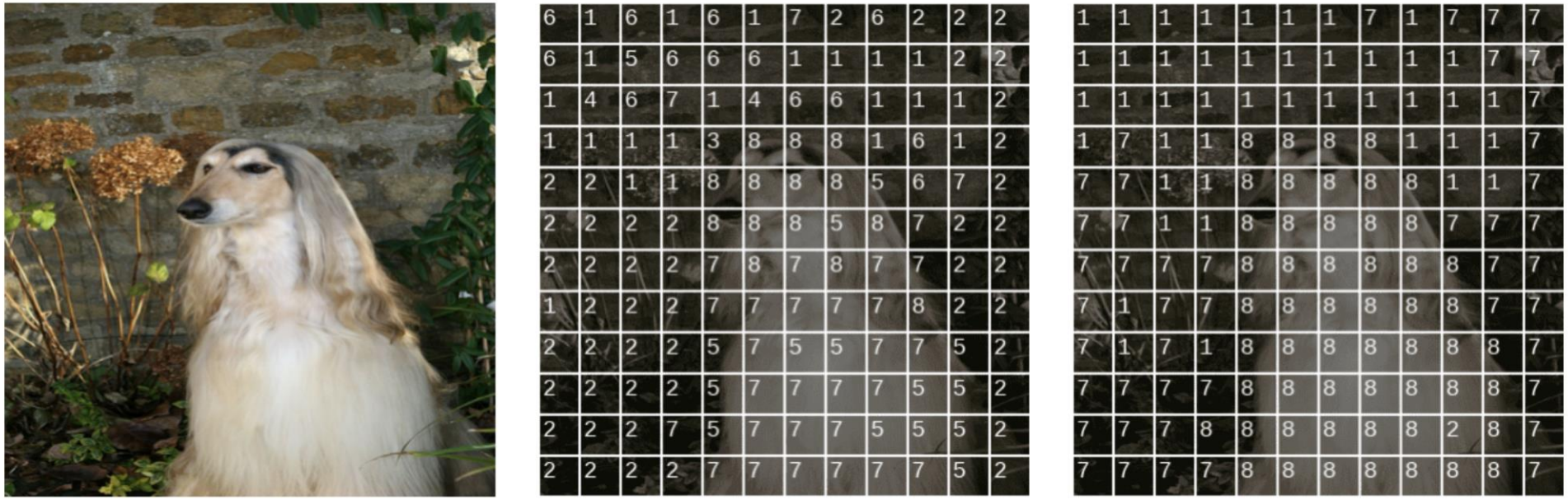
We evaluated the change rate of the selected experts before and after data augmentation.

	Top-1	Top-2	Top-2 (order-agnostic)
V-MoE-S	63.04%	34.77%	45.44%
V-MoE-S w/ PRC	75.78%	48.54%	58.96%

By applying PRC, the router becomes robust to small changes in the input.

Analyses | Visualization of Router

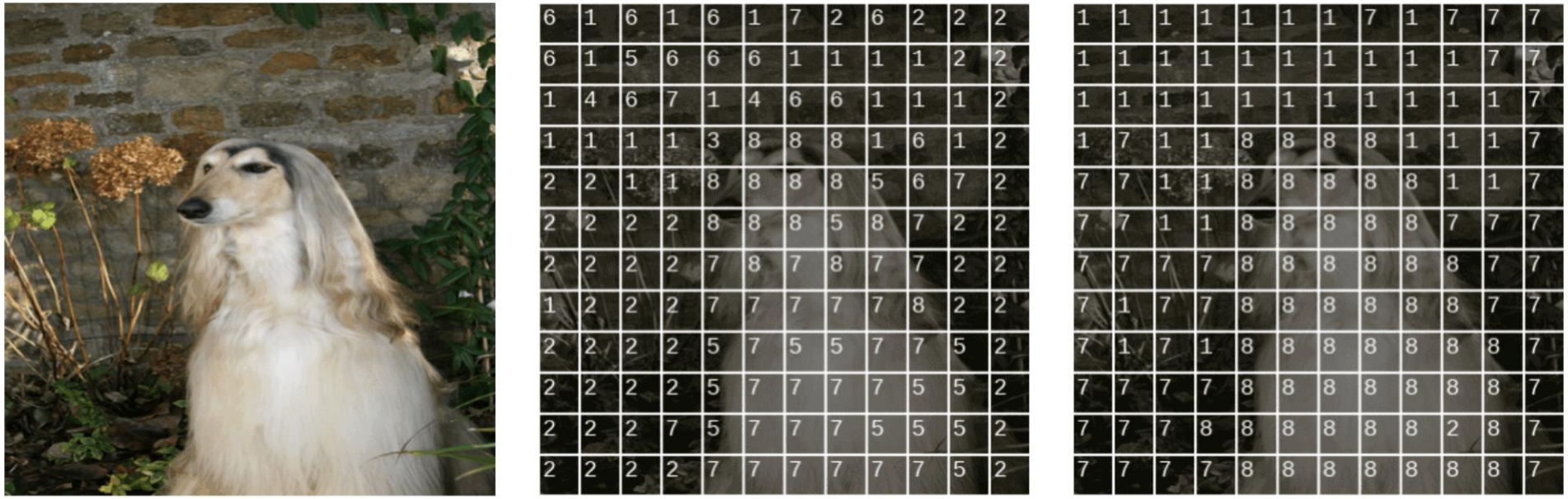
We visualized the router as Gaussian noise was gradually applied.



The numbers in the image are the selected experts.

Analyses | Visualization of Router

We visualized the router as Gaussian noise was gradually applied.

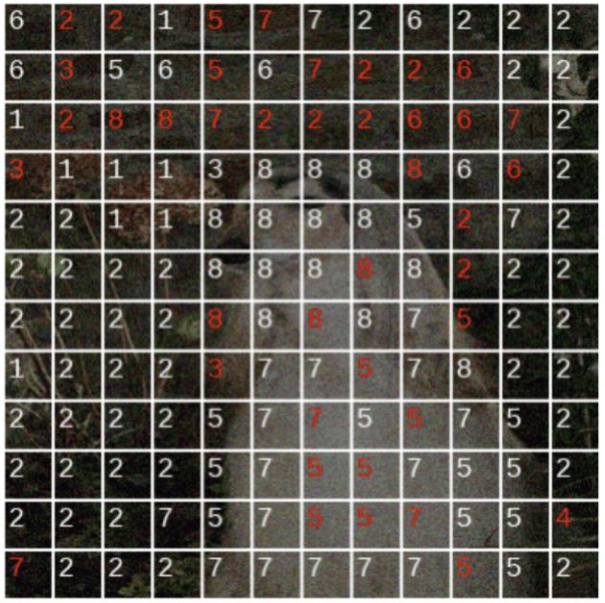


The numbers in the image are the selected experts.

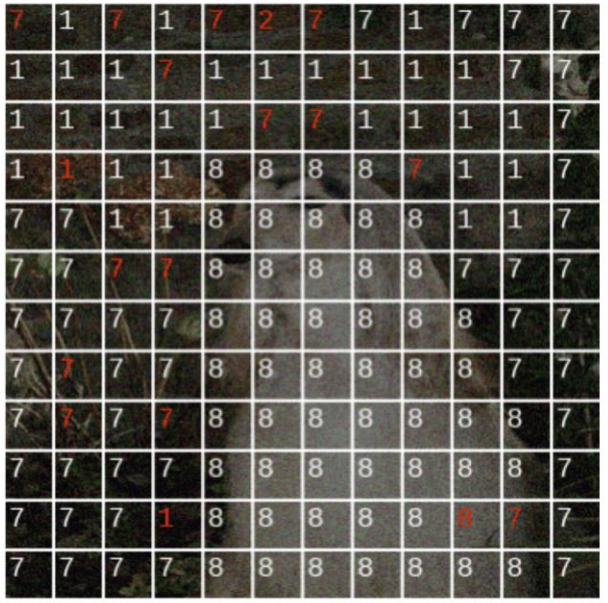
Analyses | Visualization of Router

We visualized the router as Gaussian noise was gradually applied.

Noised Image



V-MoE



V-MoE w/ PRC

By applying PRC, the router becomes robust to small changes in the input.

Summary

Purpose	Making the router robust to input perturbations and alleviating the problem that MoE output changes discretely.
Method	We propose PRC, which brings the router outputs of two data-augmented images closer together.
Evaluation	By applying PRC, we improved image classification accuracy and enhanced the router's robustness to the input perturbations. Additionally, we successfully reduced the number of experts used without a deterioration in accuracy.