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

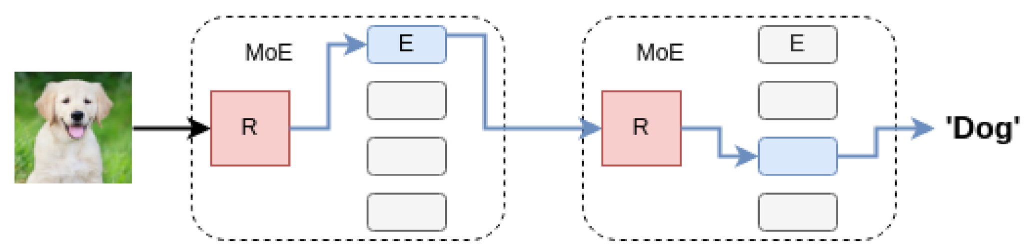
Masahiro Kada[1], Ryota Yoshihashi[1], Satoshi Ikehata[1, 3], Rei Kawakami[1], Ikuro Sato[1,2]

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

1. Introduction

Background: Sparse Mixture-of-Experts (MoE)

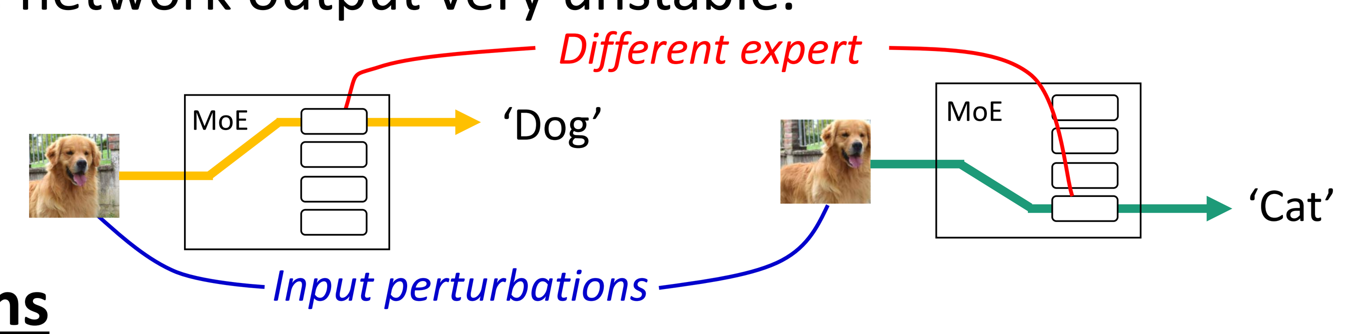
MoE is a neural network that includes **experts** specialized in certain inputs and a **router** that selects an expert or a few experts.



[N. Shazeer+, ICLR2017]
[D. Dai+, arXiv2401.06066]
[C. Riquelme+, NeurIPS2021]

Issue of Sparse MoE

MoE's output may change discontinuously due to input perturbations, making the network output very unstable.



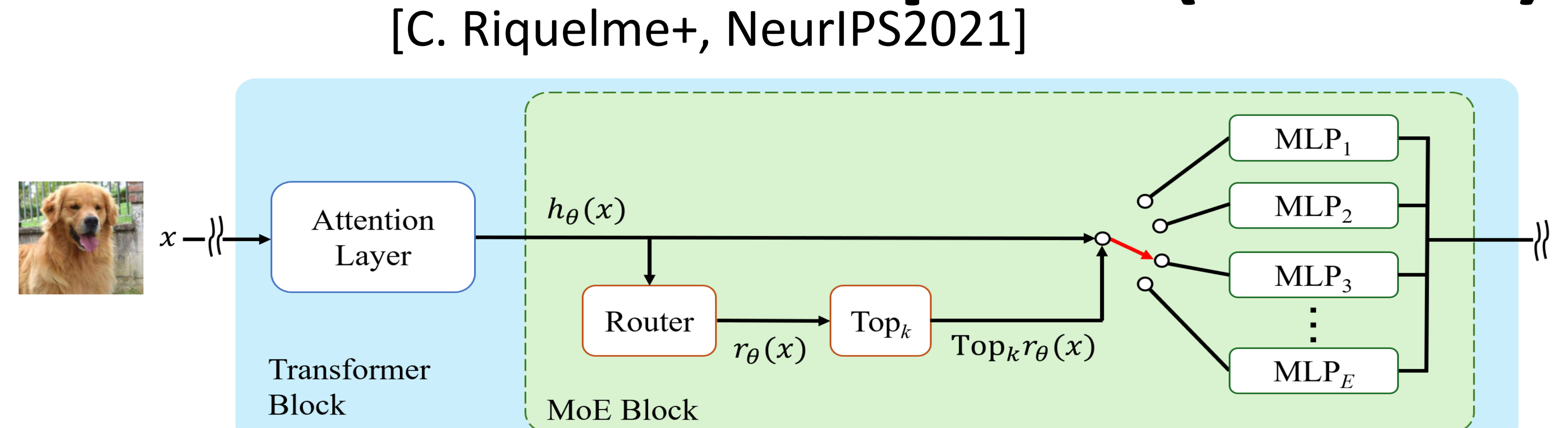
Contributions

- To address this issue, we propose **Pairwise Router Consistency (PRC)** for regularizing the router so that its output becomes robust under input data augmentation.
- We demonstrate that PRC yields more consistent expert selections under input data augmentation.
- Sparse MoEs trained with PRC achieve higher image classification accuracies on ImageNet-1K and CIFAR-10/100 datasets.

2. Vision Mixture of Experts (V-MoE)

[C. Riquelme+, NeurIPS2021]

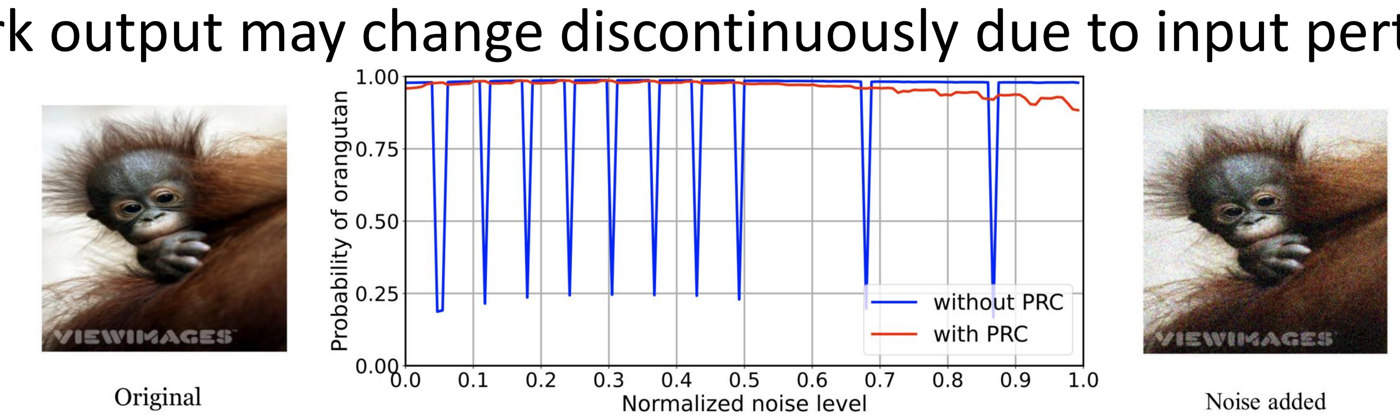
Model Design



- MLPs in the ViT transformer blocks are replaced with MoE blocks.
  - Expert selection is performed for each patch token.
- Router has learnable parameters with softmax output.
- Top-k of the router output is/are multiplied to corresponding expert output.

Issue of V-MoE

Network output may change discontinuously due to input perturbations.



3. Proposed Method

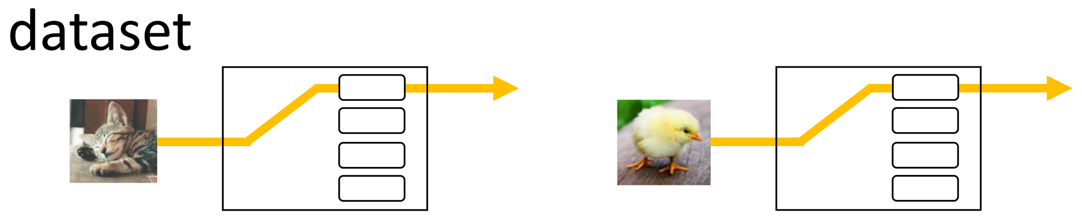
Pairwise Router Consistency (PRC)

- PRC regularization loss is designed to penalize a router that is sensitive to data augmentation, biased toward particular experts, or ambiguous in expert selection.
- We add the following PRC regularization loss term for each of the routers in an MoE.

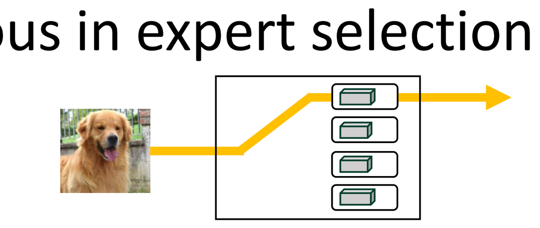
Routers penalized by the PRC:

1. Sensitive to data augmentation

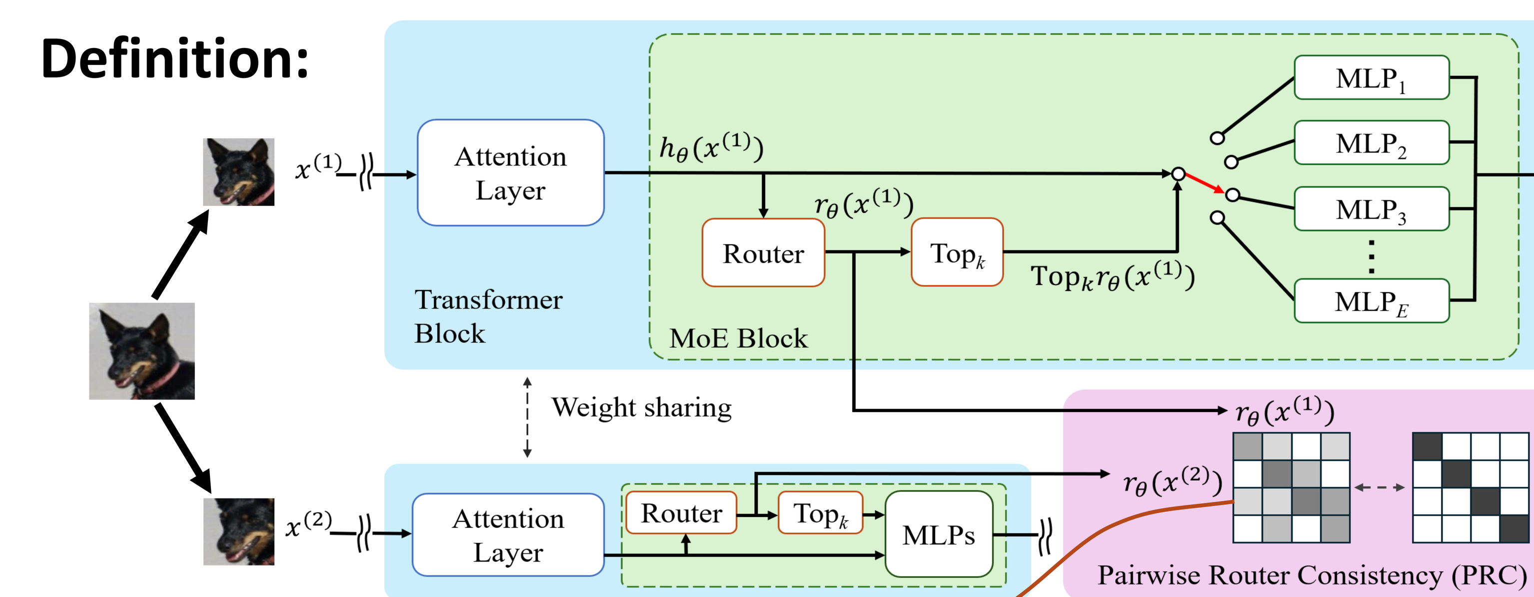
2. Biased toward particular experts within a dataset



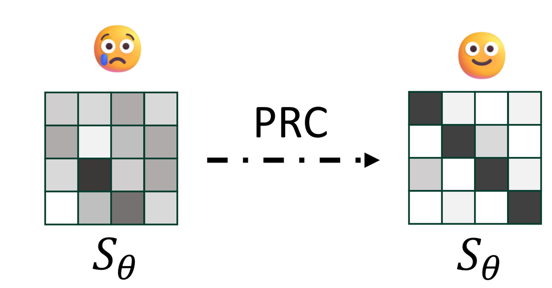
3. Ambiguous in expert selection



Definition:



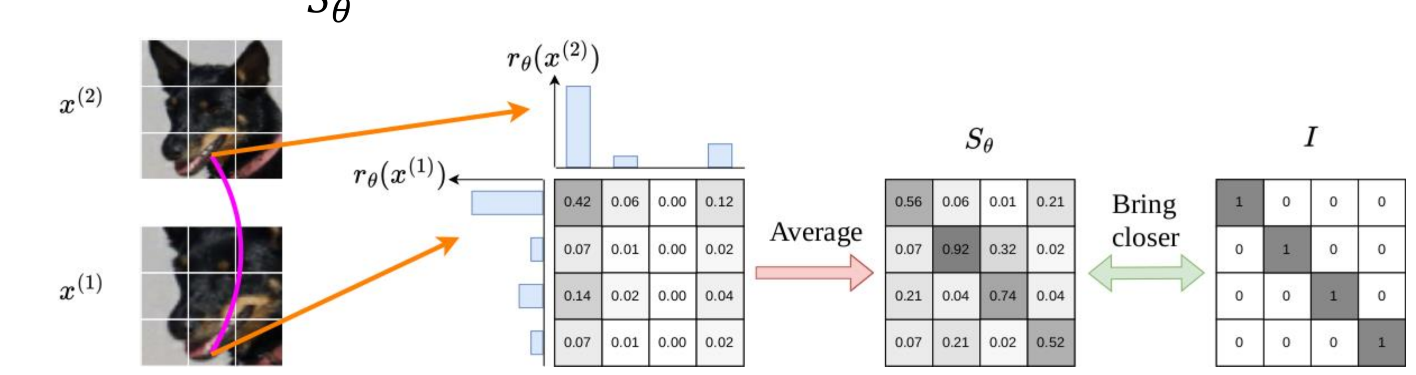
$$L_\theta^{PRC} = \| S_\theta - I \|_2^2, \quad S_\theta = C \sum_{x \in X} r_\theta(x^{(1)}) r_\theta(x^{(2)})^T$$



- Robust to data augmentation.
- Different experts are equally utilized.
- Returns an output close to one-hot.

Implementation to V-MoE

When geometrical deformation is adopted in data augmentation, an extra processing is required to identify image patches between two data-augmented samples.



4. Evaluation

Quantitative Result

- PRC empirically improves classification accuracy on ImageNet-1K, CIFAR-10/100 datasets, compared to the baseline method.
- Surprisingly, PRC with k=1 outperforms non-PRC with k=2. (k is the # of experts selected.)

Table. Image classification accuracy. k is the # of experts selected.

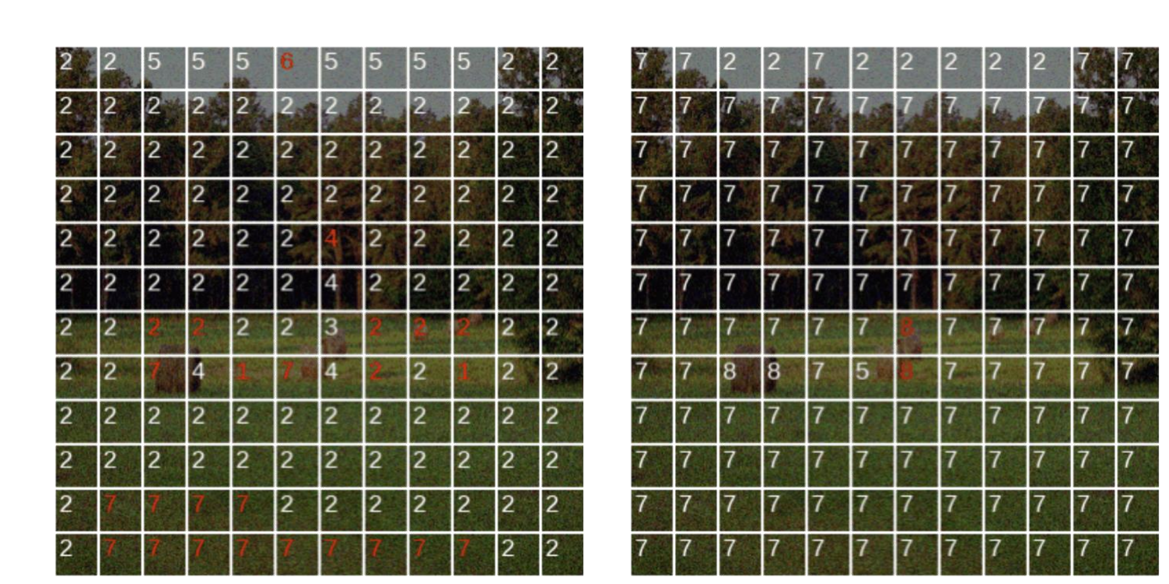
	k	ImageNet-1K	CIFAR-10	CIFAR-100	Flowers
V-MoE-S	2	75.84%	95.20%	81.38%	89.30%
V-MoE-S w/ PRC	2	**76.27%**	**95.36%**	**82.27%**	**90.18%**
V-MoE-S	1	75.23%	94.81%	81.18%	90.21%
V-MoE-S w/ PRC	1	**75.92%**	**95.12%**	**82.12%**	**91.24%**

Analysis of Router Output

- With PRC, the argmax of the router's output is better preserved under input data augmentation.
- Visualization of the router's output also indicates that expert selection frequently changes without PRC even for a relatively simple natural image.

Table. Rate of the # of expert blocks with consistent expert selection under 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%**



V-MoE **V-MoE w/ PRC**

Figure. Router output when Gaussian noise is added to an image. Routing changes are shown in red.

This work was supported by JSPS KAKENHI Grant Number JP22H03642 and DENSO IT LAB Recognition and Learning Algorithm Collaborative Research Chair (Science Tokyo).