Path Representation Learning of Mixture of Experts Based on Contrastive Learning

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

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

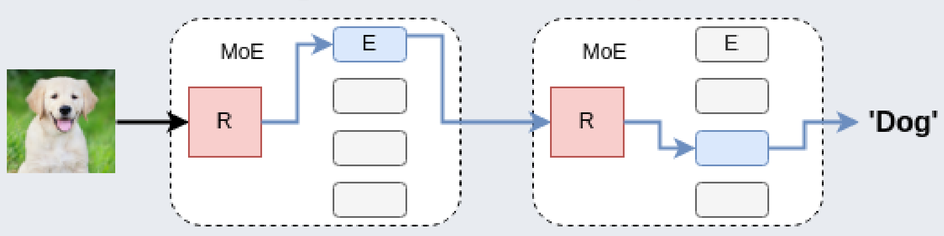
東京工業大学 Tokyo Institute of Technology

Recognition and Learning Algorithm Laboratory

DENSO IT LAB

1. Deep Mixture of Experts (MoE)

What A type of new NN architectures containing Router and Expert blocks.



Advantage Scalability: Can add experts while keeping inference time constant.

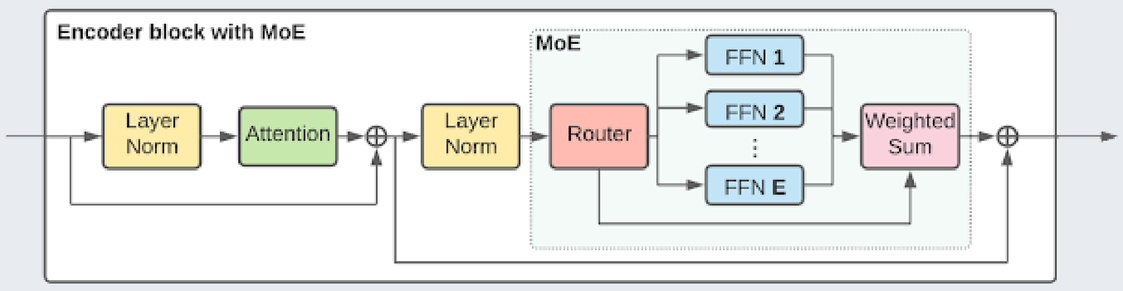
Dis-advantage Vulnerability: Output can change abruptly due to the change in feedforward path even for small input deformation.

2. Contribution

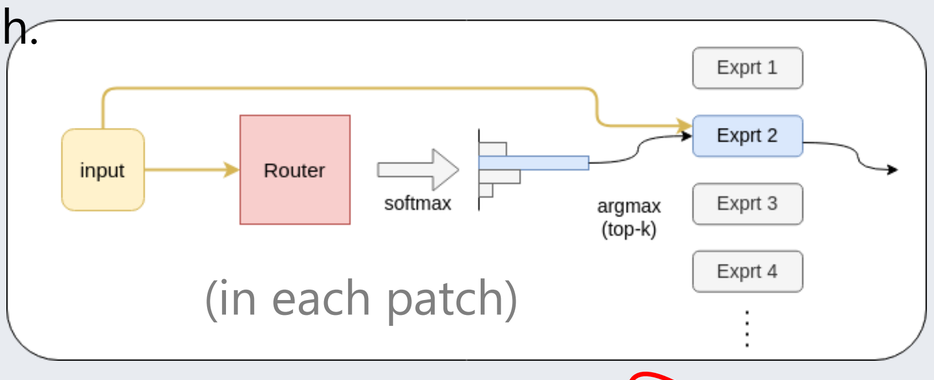
- Proposed Path Representation Learning that aims to robustify router output (`path') under input deformations.
- Confirmed that local path representations of transformer-based MoE become more insensitive to input deformations.
- Improved the performance over the previous method.

3. Previous Method

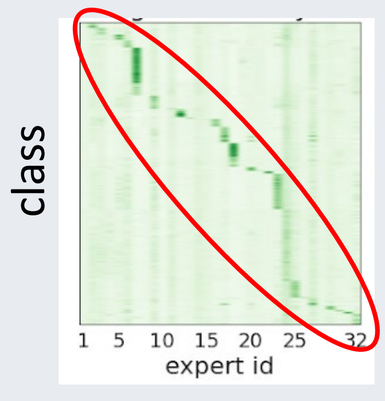
[C. Riquelme+, NeurIPS2021] proposed Vision Transformer-based MoE.



The Router selects an expert in a patch-wise manner by taking the arg-max of softmax in each patch.

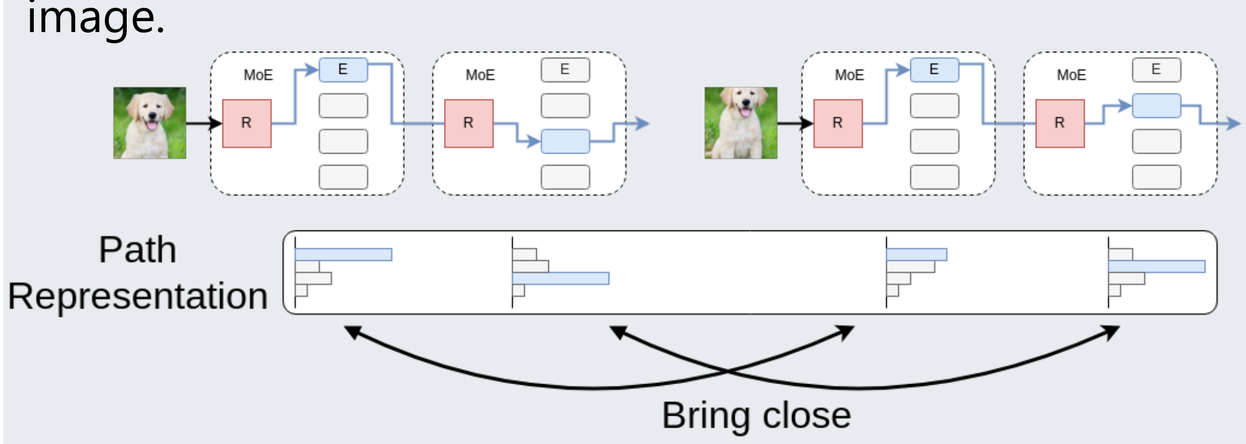


Observation
Object labels tend to be *roughly* associated with paticular experts after supervised training.

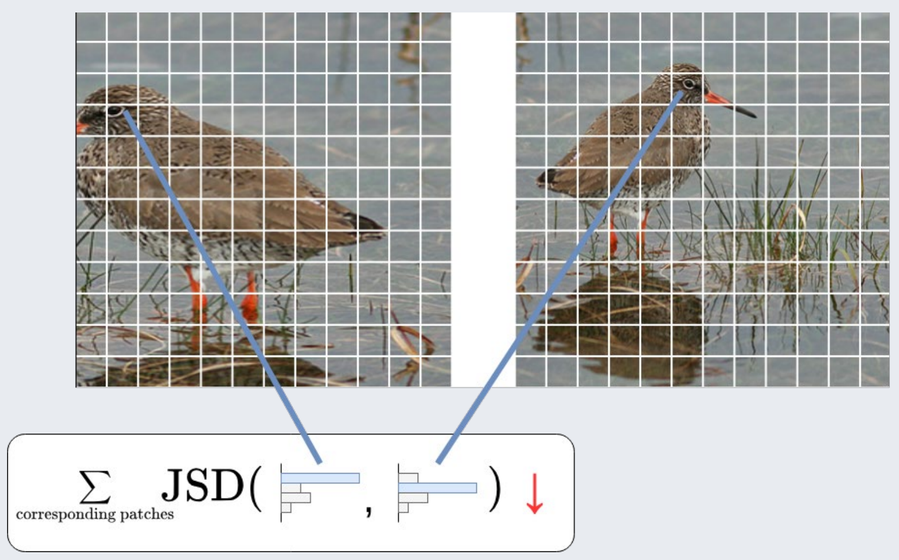


3. Proposed Method

Basic idea
Making the router outputs (paths) closer to each other between a pair of images augmented from the same image.



Applied to Vision Transformer
Making the router outputs (paths) closer in each pair of corresponding image patches.



JSD: Jensen Shannon Divergence

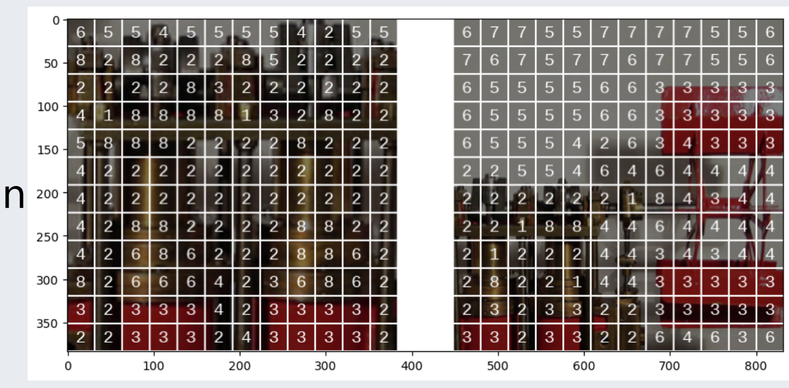
$$\mathrm{JSD}(p,q) = \frac{1}{2}\sum\left(p\log\frac{p}{M} + q\log\frac{q}{M}\right) \quad \text{, where } M = \frac{p+q}{2}$$

4. Preliminary Experiment

Qualitative results

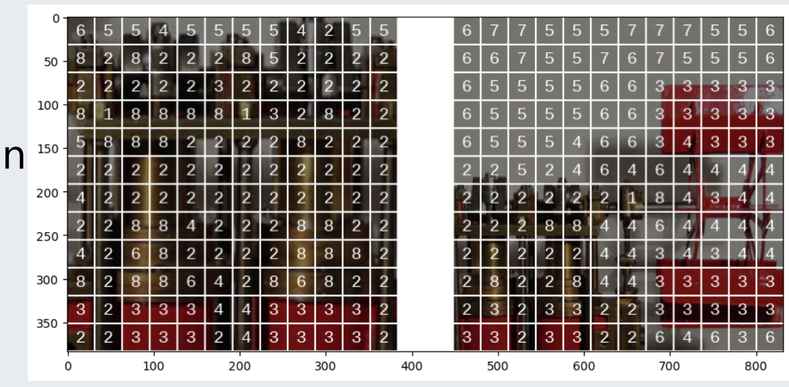
Previous method

Inconsistencies of expert selections in corresponding patches



Our method

More consistent expert selections in corresponding patches



Classification Accuracy Dataset: ImageNet-1K

Outperformed [C. Riquelme+, NeurIPS2021] slightly.

	Accuracy
Previous method	75.84%
Ours	**75.91%**

This work is supported by DENSO IT LAB Recognition and Learning Algorithm Collaborative Research Chair (Tokyo Tech.).