ルータ蒸留を用いたMixture of Expertsの高並列化

加太 将弘[1], 吉橋 亮太[1], 池畑 諭[2,3], 川上 玲[1], 佐藤 育郎[1,2]

[1]東京科学大学, [2]デンソーITラボラトリ, [3]国立情報学研究所

1. Sparse Mixture of Experts (MoE)

背景 経験上パラメータ数の拡大に従いモデル性能は向上(スケーリング則)
[arXiv2001.08361]

パラメータの増加に伴い,
計算コストが増大

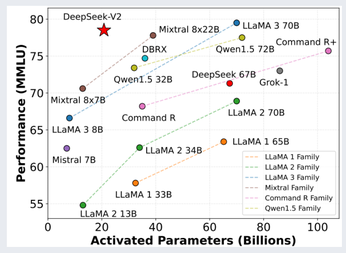
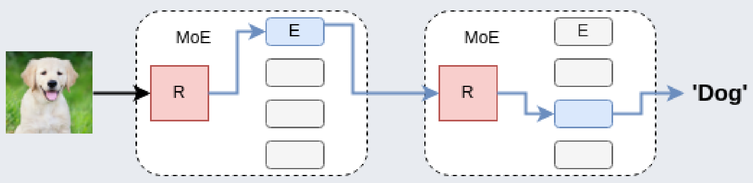


図. パラメータと性能の関係
[arXiv2405.04434]より引用

概要 特定のクラスの処理に特化した
エキスパートと少数のエキスパート選択を行うルータ
によって構成される分岐構造を有する深層モデル



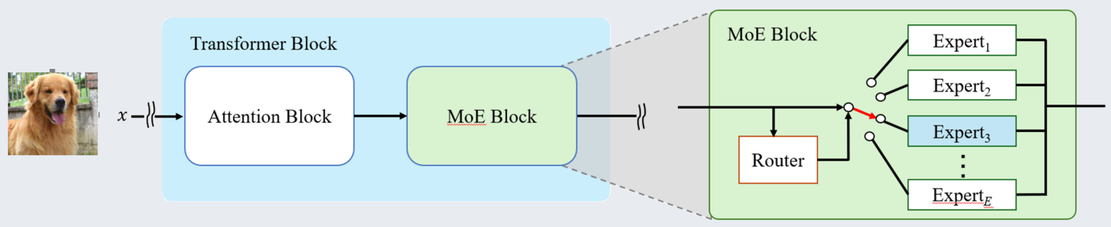
計算コストを一定に保ちつつモデルサイズの拡大可能で，スケーリング則を超える可能性を持つ手法として注目を集めている

DeepSeek[arXiv2401.06066] など最新のLLM/VLMで採用

2. 視覚モデルにおける従来法

Vision Mixture of Experts (V-MoE)[C. Riquelme+, NeurIPS2021]

Vision Transformer[A. Dosovistskiy+, ICLR2021]のMLP層をMoE化



ルータはSoftmaxとして定義され，Top-K個の要素を選択

課題
- 入力の摂動に対して出力が離散的に変化
  PRC[M. Kada+, ICASSP2025]にて解決
- 経験的にエキスパート数を増やすと汎化が劣化に転じる

3. 提案法: ルータ蒸留

着想 離散化を伴うルーティングは勾配計算ができず適切なエキスパート割り当てが困難
特にエキスパートの増加によりその問題が顕著に

↓

非MoEの同じアーキテクチャの教師モデルを知識蒸留の形でルータ学習に活用すれば，教師の知識を反映した安定かつ意味的整合性のあるエキスパート選択が可能になるのではないか？

手法 1. 教師ルータの学習

- 非MoEの教師モデルの特徴を用いてルータを構築
- エキスパート選択頻度を均等化するロードバランシング損失を適用して最適化

教師モデルのパラメータθ_Tを次式で最適化

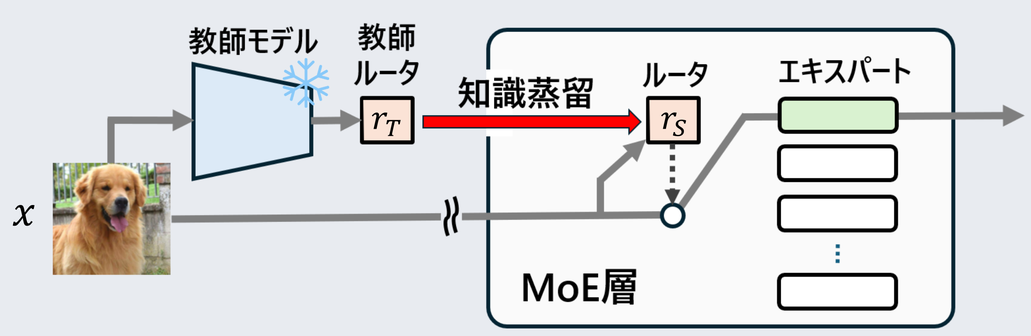
$$\theta_T^* = \mathrm{argmin}_{\theta_T}\ L_{\mathrm{load}}$$

$$L_{load} = \mathrm{Var}\left(\sum_x r_T\right)$$

※ただし， 教師モデルのbackboneは更新しない

2. ルータの知識蒸留

- 学習済みの教師ルータ出力を生徒ルータへ知識蒸留



生徒モデルのパラメータθ_Tを次式で最適化

$$\theta^* = \mathrm{argmin}_{\theta}\ L_{\mathrm{task}} + L_{\mathrm{router\_distill}}$$

$$L_{\mathrm{router\_distill}} = \mathrm{KL}(r_T \parallel r_S)$$

4. 評価

実験設定 データセット: ImageNet-1K
生徒モデル: DeiT[T. Hugo+, ICML2021] (Tiny / Small)をMoE化
教師モデル: ImageNet-21Kで学習したDeiT Small

提案法の有効性

ImageNet-1K上での実験により，MoE化したDeiTに提案法を適用することで，Smallで+0.85%，Tinyで+0.24%の精度向上を確認

エキスパート数の影響

エキスパート数を増加させた場合にも一定の効果は見られたが，大幅な精度改善には繋がらなかった

表1. ImageNet-1Kの分類精度

	エキスパート数	ImageNet-1K精度
DeiT-MoE-S	8	82.22%
DeiT-MoE-S w/ Router Distill	8	**83.07%**
DeiT-MoE-Ti	8	76.55%
DeiT-MoE-Ti w/ Router Distill	8	**76.79%**

表2. エキスパート数を増やした際のImageNet-1Kの分類精度

	エキスパート数	ImageNet-1K精度
DeiT-MoE-S	32	82.27%
DeiT-MoE-S w/ Router Distill	32	**83.19%**

5. まとめ

提案法により，教師モデルの知識を活用したルーティングが可能となり，画像分類精度向上を確認

6. 今後の展望

- 高並列化に関してさらなる改善が必要
- 今後，自然言語処理モデルを用いて比較評価

本研究は東京科学大学「DENSO IT LAB 認識・制御・学習アルゴリズム共同研究講座」における研究プロジェクト「機械学習の実応用に向けた品質基盤開発」の成果である